International Journal of Fog Computing
Volume 2 • Issue 1 • January-June 2019

Overview of Big Data-Intensive Storage and its Technologies for Cloud and Fog Computing

Richard S. Segall, Arkansas State University, Jonesboro, USA

Jeffrey S Cook, Independent Researcher, Paragould, USA

Gao Niu, Bryant University, Smithfield, USA

ABSTRACT

Computing systems are becoming increasingly data-intensive because of the explosion of data and the needs for processing the data, and subsequently storage management is critical to application performance in such data-intensive computing systems. However, if existing resource management frameworks in these systems lack the support for storage management, this would cause unpredictable performance degradation when applications are under input/output (I/O) contention. Storage management of data-intensive systems is a challenge. Big Data plays a most major role in storage systems for data-intensive computing. This article deals with these difficulties along with discussion of High Performance Computing (HPC) systems, background for storage systems for data-intensive applications, storage patterns and storage mechanisms for Big Data, the Top 10 Cloud Storage Systems for data-intensive computing in today's world, and the interface between Big Data Intensive Storage and Cloud/Fog Computing. Big Data storage and its server statistics and usage distributions for the Top 500 Supercomputers in the world are also presented graphically and discussed as data-intensive storage components that can be interfaced with Fog-to-cloud interactions and enabling protocols.

KEYWORDS

Cloud Storage, Data-Intensive, Fog-To-Cloud Computing, High Performance Computing (HPC), Key-Valued, Message Passing Interface (MPI), Storage Systems, Supercomputers

INTRODUCTION

Data-intensive computing systems have penetrated every aspect of people's lives. Behind it is the scientific and commercial processing of massive data impacting the decision makings in companies, academics, governments, social cites, and personal lives.

There are two types of data-intensive computing systems that continue to co-exist in the modern computing environment:

1. High Performance Computing (HPC) systems, consisting of tightly coupled computer nodes and storage nodes that are used to execute task parallelism for scientific purposes like weather forecasting, physics simulation, and the likes. (Rouse, 2017b).

DOI: 10.4018/IJFC.2019010104

2. Message Passing Interface (MPI) is an example of a computing framework on HPC systems. Big Data systems, comprised of more loosely coupled nodes, are used to execute data parallelism for tasks such as sorting, data mining, machine learning, etc. MapReduce is an example of a computing framework on Big Data systems. ((Barney, 2017) (Rouse, M. (2017c)).

Both HPC systems and Big Data systems that are deployed for multiple users and applications to share the computing resources so that 1) the resource utilization is high, driving down the usage cost per application/user, and the users get better responsiveness of application execution; 2) the data set is reused without extra overhead to move around performing redundant Input/Outputs (I/O) and users can also save space.

As the computing needs continue to grow in data-intensive computing systems, the shared usage model results in a highly resourceful competing environment. For example, Amazon, Apple and eBay provides HPC and Big Data as cloud services. Hadoop version 2, YARN (Yet Another Resource Negotiator), that is one of the key features in the second-generation Hadoop 2 version of the Apache Software Foundation's open source distributed processing framework. Originally described by Apache as a redesigned resource manager. YARN is now characterized as a large-scale, distributed operating system for Big Data applications which provides a scheduler to incorporate both MapReduce and MPI jobs. (Rouse, 2017a).

As the number of concurrent data-intensive applications and the amount of data increase, application I/O's start to saturate the storage and interfere with each other, and storage systems become the bottleneck to application performance. Both HPC and Big Data systems I/O amplification adds to the I/O contention in the storage systems. To counter failures in these distributed systems, HPC systems employ defensive I/O's such as check pointing to restart an application from where it fails, and Big Data systems replicate persistent data by a factor of k, which grows with the scale of the storage system. Both mechanisms aggravate the I/O contention on the storage. The storage systems can be scaled-out, but the compute to storage node ratio is still high, rendering the storage subsystem a highly contended component (Xu, 2016). Therefore, the lack of I/O performance isolation in the data-intensive computing systems causes severe storage interference which compromises the performance target set by other resource managers proposed or implemented in a large body of works. Failure to provide applications with guaranteed performance has consequences. Data-intensive applications must complete in bounded time so as to get meaningful results. For example, weather forecast data is much less useful when the forecasted time has passed. Paid user in a Big Data system also require a predictable runtime even though the job is not time sensitive, and the provider may get penalized in revenues if jobs fail to complete in a timely manner. (Xu, 2016).

This chapter addresses the problems stated above for data-intensive computing systems. It provides different approaches for both HPC storage systems and Big Data storage systems because their differences in principles, architecture, and usage pose distinct challenges. Before studying these systems and addressing their respective problems separately, the discussion of the differences between these two types of systems is established here. (Xu, 2016).

HPC systems are strongly coupled distributed systems, connected by expensive hardware and network links (e.g. InfiniBand [inf]). The application execution principle focuses on *task parallelism*, and thus both its parallel compute processes and I/O requests are tightly coupled and must be executed *together*. This means a failure of any node results in the failure of the entire application. This is also why the check pointing I/O's are major sources of I/O's when running such applications, as the periodical save of application progress constitutes much higher amount of data than its original input and final output. (Xu, 2016).

The most widely used programming framework for HPC systems is Message Passing Interface (MPI) that is also discussed in this article.

FOG COMPUTING

Fog computing is a term created by Cisco that refers to extending cloud computing to the edge of an enterprise's network. Also known as Edge Computing or fogging, Fog Computing facilitates the operation of compute, storage, and networking services between end devices and cloud computing data centers.

Balakrishnan, Venkatesh & Raj (2018) presented an introduction, architecture, analytics, and platforms of Fog Computing, and Segall & Niu (2018) presented a presented a preceding article in this journal *International Journal of Fog Computing* (IJFC) on an overview of Big Data and its visualizations with Fog Computing.

While Edge Computing typically refers to the location where services are instantiated, Fog Computing implies distribution of the communication, computation, and storage resources and services on or close to devices and systems in the control of end-users. (Ostberg et al. (2017), Chiang & Zhang (2016)). Fog Computing is a medium weight and intermediate level of computing power (Pierra (2017)).

Ostberg et al. (2017) discussed reliable capacity provisioning for distributed cloud/edge/fog computing applications. According to Alageswaran & Amali (2018) states that it is generally considers that Fog Computing as the appropriate platform for many applications and especially suited for the Internet of Things (IoT). Alageswaran & Amali (2018) also state the different characteristics of Fog Computing as consisting of: edge location, location awareness, and low latency; geographical distribution, support for mobility, real-time interactions, heterogeneity of fog nodes being deployed in a wide variety of environments, and interoperability of fog components in order to give a wide range of services such as data-streaming.

Bhatt & Bhensdadia (2018) discussed Fog Computing in IoT noting that the IoT utilizes various platforms like GoogleCloud, Amazon, and GENI and that Fog Computing/Cloudlets/Edge Computing acts as a connection between devices and cloud computing. Fog computing also provides an intelligent platform to manage the distributed and real-time nature of emerging applications and infrastructures. Perera et al. (2017) presented a survey of Fog Computing for sustainable smart cities.

Fog Computing offers services that deliver higher delay performance due to nearness to the end-users compared to the cloud data centers Cloud has large machines; storage and communication abilities compared to the fog and may interface with High-performance computing (HPC) systems as described in next section (Bhatt & Bhensdadia (2018).

According to Cisco (2015) report, Fog Computing extends the cloud to be closer to the things or devices that produce data and act on IoT data. These devices called "fog nodes" can be can be deployed anywhere with a network connection. Cisco (2015) states that any device with computing, storage, and network connectivity can be fog node.

Belli et al. (2018) discusses an in-depth study of a proposed scalable Big Stream cloud architecture for the IoT that reverses the traditional "Big Data" paradigm where real-time constraints are not considered. Alageswaran & Amili (2018) discussed the evolution of Fog Computing and its role in IoT applications. Alonso-Monsalve, Garcia-Carballeria, & Calderon (2017) discussed the concept of storage with Fog Computing and its role in IoT applications.

However, according to Alageswaran & Amali (2018) not every fog device may be able to provide data storage service over long time, and thus require storage on cloud server or data-intensive storage systems in a fog environment level as discussed later in this article.

FOG-TO-CLOUD COMPUTING

According to Chiang & Zhang (2016) one of the reasons that fog is an emerging era in relation to data storage is that it can carry out a substantial amount of data storage at or near the end-user rather than storing data in remote data centers.

However, one needs to understand that the Fog-Cloud interface supports Fog-Cloud collaborations to provide end-to-end services. Ahuja & Deval (2018) discussed Fog-to-Cloud Computing using Internet-of-Things (IoT) as platforms. Raj & Raman (2018) published a handbook of research on cloud and Fog computing infrastructures for data science. According to Chiang & Zhang (2016), the Fog-Cloud interface supports functions such as the following:

1. Fog to be managed from the Cloud
2. Fog and Cloud to send data to each other
3. Cloud distribute services onto Fog
4. Cloud services to be provided to Fog
5. Cloud services to be provided through Fog-to-Things and end users
6. Fog services to be provided to Cloud
7. Fog and Cloud to collaborate with each other to deliver end-to-end services

Hence, the Fog-to-Cloud computing can be controlled by the Fog-Cloud interface such as that have High-Performance Computing Systems as discussed in the following section and elsewhere in this paper. High-performance computers can be located either as a Fog node or in a remote location with with data centers. Hence, users of Fog-to-Cloud Computing need to understand the potentials that exist of cloud services able to provide to Fog, and that Fog and Cloud are able to send data to each other that may be in the category of Big Data.

This article discusses Cloud storage systems for data-intensive computing that can be interfaced with Fog to collaborate with each other to deliver end-to-end services. This chapter also discusses storage mechanisms, patterns and technology for Big Data that may interface with Fog-to-Things and end users. As discussed in Segall (2017a, 2017b), tablets and mobile devices can be used for visual analytics of Big Data in bioinformatics, as well as technologies for teaching Big Data Analytics in the classroom.

HIGH PERFORMANCE COMPUTING (HPC) SYSTEMS

High-performance computing (HPC) is the use of parallel processing which is the processing of program instructions by dividing them among multiple processors with the objective of running a program in less time. In the earliest computers, only one program ran at a time. A computation-intensive program that took one hour to run and a tape copying program that took one hour to run would take a total of two hours to run. An early form of parallel processing allowed the interleaved execution of both programs together. The computer would start an I/O operation, and while it was waiting for the operation to complete, it would execute the processor-intensive program. The total execution time for the two jobs would be a little over one hour for running advanced application programs efficiently, reliably and quickly. (Rouse, 2017b)

The term HPC applies especially to systems that function above a teraflop or 10 to the 12^{th} power floating-point operations per second. The term HPC is occasionally used as a synonym for supercomputing, although technically a supercomputer is a system that performs at or near the currently highest operational rate for computers. Some supercomputers work at more than a petaflop or 10 to the 15th power floating-point operations per second. (Rouse, 2017b)

Previous work on HPC by the authors of this article include that of Segall (2013), Segall (2016a), Segall (2016b), Segall (2015), Segall, Cook & Zhang (2015), and Segall & Gupta (2015).

The most common users of HPC systems are scientific researchers, engineers and academic institutions. Some government agencies, particularly the military, also rely on HPC for complex applications. High-performance systems often use custom-made components in addition to so-called commodity components. As demand for processing power and speed grows, HPC will likely interest

businesses of all sizes, particularly for transaction processing and data warehouses. An occasional techno-fiends might use an HPC system to satisfy an exceptional desire for advanced technology. (WhoIsHostingThis.com, 2017) High-performance computing and data mining in bioinformatics was discussed in Segall (2016).

Typically, the problems under consideration cannot be solved on a commodity computer within a reasonable amount of time (too many operations are required) or the execution is impossible, due to limited available resources (too much data is required). HPC is the approach to overcome these limitations by using specialized or high-end hardware or by accumulating computational power from several units. The corresponding distribution of data and operations across several units requires the concept of parallelization. When it comes to hardware setups, there are two types that are commonly used: (COMSUL, Inc., 2017)

1. Shared memory machines
2. Distributed memory clusters

Shared Memory Machines

In shared memory machines, Random-Access Memory (RAM) can be accessed by all of the processing units. (COMSUL, Inc., 2017)

Distributed Memory Machines

Meanwhile, in distributed memory clusters, the memory is inaccessible between different processing units, or nodes. When using a distributed memory setup, there must be a network interconnect to send messages between the processing units (or to use other communication mechanisms), since they do not have access to the same memory space. (COMSUL, Inc., 2017)

MODERN HIGH-PERFORMANCE COMPUTING (HPC) HYBRID SYSTEMS

Modern HPC systems are often a hybrid implementation of both concepts, as some units share a common memory space and some do not. HPC is primarily used for two reasons. First, due to the increased number of Central Processing Units (CPUs) and nodes, more computational power is available. Greater computational power enables specific models to be computed faster, since more operations can be performed per time unit. This is known as the speedup. The speedup is defined as the ratio between the execution time on the parallel system and the execution time on the serial system. The upper limit of the speedup depends on how well the model can be parallelized. (COMSUL, Inc., 2017)

Consider, for example, a fixed-size computation where 5% of the code is able to be parallelized. In this case, there is a theoretical maximum speedup of 2. If the code can be parallelized to 95 it is possible to reach a theoretical maximum speedup of 20. For a fully parallelized code, there is no theoretical maximum limit when adding more computational units to a system. Amdahl's Law explains such a phenomenon. (COMSUL, Inc., 2017)

Amdahl's Law is a formula used to find the maximum improvement by improving a particular part of a system. In parallel computing, Amdahl's Law is mainly used to predict the theoretical maximum speed up for program processing using multiple processors. (Techopedia, 2017)

STORAGE SYSTEMS FOR DATA-INTENSIVE APPLICATIONS

Background

According to the US Department Energy (DOE) Advanced Scientific Computing Advisory Committee (ASCAC) Report of 2013 (Chen et al., 2013), storage systems for data-intensive applications are

pervasive and indispensable subsystems in current data-intensive computing systems. In High Performance Computing (HPC) systems, Parallel File Systems (PFS) are used to aggregate the throughput from multiple storage nodes to serve the concurrent, parallel I/O requests from the user applications. In Big Data computing systems, Distributed File Systems (DFS) are used not only for the parallelism of jobs, but also the data-locality and replication of data.

Figure 1 (Chen, J. et al, 2013) from the DOE ASCAC Report compares "compute intensive architecture" versus "data intensive architecture" in the 2017 time frame for many characteristics such as "On-node-storage", "In-Rack Storage", "Global Shared Disk, and "Off-System Network".

Table 1 (Chen et al, 2013) from the DOE ASCAC Report presents data-generation requirements for different domains. Table 1 provides four different scenarios for data generation phase and compares the transactional processing requirements, storage or post processing, sharing and distribution, and visualization. For example, computational biologists may perform DNA sequencing by using Next Generation Sequencing (NGS) machines produced by companies such as Illumina, Roche, and Applied Biosystems. Winn et al. (2012) discusses data-intensive computing in biology and the storage systems used.

The ICON Group International (2017) published a 266-page report titled *"The 2018-2023 World Outlook for Big Data Storage"* This study covers the world outlook for Big Data storage across more than 190 countries. This study estimates for the worldwide latent demand, or the Potential Industry Earnings (P.I.E.), for Big Data storage. It also shows how the P.I.E. is divided across the world's regional and national markets. In order to make estimates over time, a multi-stage methodology is employed that is often taught in courses on international strategic planning at graduate schools of business.

Several chapters on Big Data Storage appear in Qiang (2015) that is a collection of papers presented at the *Second International Conference on Cloud Computing and Big Data*, and entire chapter on Big Data Storage appear each in Chen, M. et al. (2014) and Sawant and Shah (2013). Grieco (2017) is an entire book on *Spark Big Data Cluster Computing in Production*.

Hoskin (2016) is another entire book on *VMware Software-Defined Storage* that presents in-depth look at VMWare next-generation storage technology that maximizes quality storage design. In Hoskin (2016) Storage-as-a-Service (STaaS) is discussed in terms of deployment through VMware technology, with insight into the provisioning of storage resources and operational management, while legacy storage and storage protocol concepts provide context and demonstrate how Virtual SAN (Storage Area Network) and Virtual Volumes are meeting traditional challenges.

Figure 1. Strawman compute-intensive vs. data-intensive computer architectures in the 2017 timeframe. Figure courtesy of NERSC. (HMC: Hybrid Memory Cube, NVRAM: Non-Volatile Random-Access Memory, SSD: Solid-State Drive, GAS: Global Address Space.) (Source: Chen, J, et al, DOE ASCAC Data Subcommittee Report, p. 8, 2013).

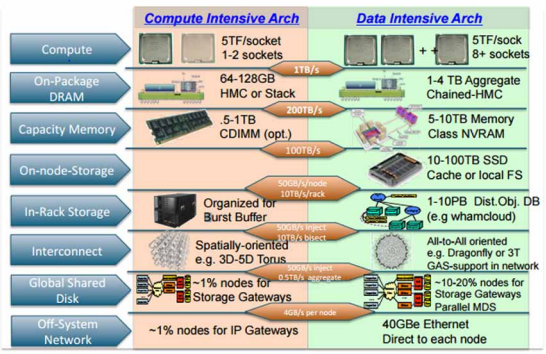


Table 1. Data-generation requirements for different domains (Source: Chen, J. et al, DOE ASCAC Data Subcommittee Report, p. 23, 2013)

Data Generation Phase (scenarios)	Overview	Transactional / in situ processing requirements	Storage for Post processing	Sharing and distribution	Visualization
1) Design Oriented Exascale Simulations e.g., combustion, CFD	Generation of data from simulations.	(1) Data reduction for post processing (2) feature detection & tracking (3) advanced analytics	Reduced data	Low (Only the producer or a few scientists may analyze data in the future)	In-situ, inter-action, feature display, uncertainty, visual debugging
2) Discovery Oriented Exascale Simulations (2) e.g., climate, cosmology	Integration of data generated from simulation and observations	1) Data reduction for post processing (2) time series (3) statistics (4) advanced analytics	(1) Raw data (2) Well organized (DB) (3) Enabled for queries	High (A large number of scientists, geographically distributed)	InfoVis and SciVis, pattern detection, correlation, clustering, ensemble vis, uncertainty
3) Large centralized instruments e.g., LHC	Data generation from large devices Extremely high rates Centralized, coordinated, controlled access	(1) HW/SW for high-rate data processing (2) derived data (3) metadata (4) Extensive queries	(1) Raw data (2) Different forms of derived data (3) Lots of distributed copies	High (A large number of scientists, geographically distributed), different sets defined by queries and other parameters	Custom user interfaces enabling query visual analysis, trajectory vis/analysis, user driven data triage/-summarization
4) Smaller distributed instruments e.g., field work, sensors, biology	Data generation from massive numbers of distributed devices, sensors	(1) Local processing and derivations (2) Local analytics (3) Integration of massive data (possibly at an exascale level system, data centers)	(1) Raw data (2) Derived data and subsets (3) Distributed copies	High (A large number of scientists, geographically distributed)	InfoVis, high dimensional vis, large-scale graphs, patterns, clustering, scalability

Li and Qiu (2014) edited a book on *Cloud Computing for Data-Intensive Applications* and presents a range of cloud computing platforms for data-intensive scientific applications. It covers systems that deliver infrastructure as a service (IaaS), including: HPC as a service (HPCaaS); virtual networks as a service; scalable and reliable storage; algorithms that manage vast cloud resources and applications runtime; and programming models that enable pragmatic programming and implementation toolkits for eScience applications.

Editors Li and Qiu (2014) includes chapters pertaining to Big Data Storage with those authored by Tudoran et al. (2017) on "Big Data Storage and Processing on Azure Clouds: Experiments on Scale and Lessons Learned", Ramakrishnan et al. (2014) on "Storage and Data Life Cycle Management in Cloud Experiments with FRIEDA", Ross et al. (2014) on Managed File Transfer as a Cloud Service," and Gao et al. (2014) on "Supporting a Social Media Observatory with Customizable Index Structure: Architecture and Performance."

Jackson et al. (2015) wrote a detailed summary on programming models and environments for cluster, cloud, and grid computing that defends Big Data. Jackson et al. (2015) portrays a survey of

how programming models that are developed for cluster clouds and grids act as a support for Big Data Analytics.

Deka (2017) edited an entire book on *NoSQL: Database for Storage and Retrieval of Data in Cloud* that includes a chapter by Reddy and Raz (2017) on Hosting and Delivering Casandra NoSQL Database via Cloud Environments.

Hu (2016) edited a book on Big Data Storage, Sharing and Security and included a chapter by Gadepally et al. (2016) on Storage and Database Management for Big Data.

Implementation of a costing model for High Performance Computing as a Service (HPCaaS) on the cloud environment was studied by Radadiya and Rohokale (2016) so that small organization can afford to have their own HPC system at low costs.

Achahbar and Abid (2015) evaluate the impact of virtualization on HPCaaS. They track HPC performance under different cloud platforms, namely KVM and VMware-ESXi, and compare it against physical clusters. Each tested cluster provided different performance trends, and concluded that the overall findings proved that the selection of virtualization technology can lead to significant improvements when handing HPCaaS (High Performance Computing as a Service).

Azeem and Sharma (2016) studied the Converged Infrastructure (CI) and Hyper Converge Infrastructure (HCI) as future of Virtual Data Centers in the cloud and provided information that seems to solve many provisioning problems and act as platforms for utility and grid computing.

Wu et al. (2017a) provided an in-depth study with comparisons and taxonomy for Big Data storage and data models that include that for the three main storage models developed during the past few decades of: (1.) Block-based storage, (2.) File-based storage and (3.) Object-based storage. Wu et al. (2017a) summarizes data storage models and indicates that according to different data models, current data storage systems can be categorized into two big families: (1.) relational-stores (SQL) and (2.) NoSQL (Not Only SQL) stores that can be classified as three main groups: (i.) Key-valued stores, (ii.) Document stores and (iii.) Extensible-Record/Column-based stores that are discussed in more depth within this chapter.

Sakr (2016) published a book on a survey of Big Data 2.0 processing systems that includes a discussion of Big Data storage systems. Sakr is also one of the co-authors of Wu et al. (2017a) discussed above, and also discusses the new generation of scalable data storage systems called NoSQL and it types.

Swami et al. (2018) wrote a complete chapter on storing and analyzing streaming data as a Big Data challenge.

Storage Patterns for Big Data

According to Sawant and Shah (2013) there are four Big Data Storage Patterns: Façade, Lean, No SQL and Polygot and each of these are briefly discussed below. The reader is referred to book authored by Sawnat and Shah (2013) for more details and illustrative figures.

Façade Pattern

The Hadoop Distributed File System (HDFS) serves as the intermittent façade (or interface that hides the complexities) for the larger traditional Data Warehouse (DW) systems. Data from the different sources can be aggregated into an HDFS before being transformed and loaded into the traditional Data Warehouse (DW) and Business Intelligence (BI) tools. Data can be stored as "Structured" data after being ingested into HDFS in the form of storage in an RDBMS (Relational Data Base Management System) or in the form of appliances such as *IBM Netezza/EMC Greenplum*, NoSQL Databases such as *Cassandra/HP Vertica/Oracle Exadata,* or simply as an in-memory cache (Sawnat & Shah, 2013).

Lean Pattern

A method to uniquely identify a dataset can be accomplished using a Lean Pattern method that creates a unique row-key, while having only a one column-family and one column. The row-key name should end with a suffix of a time-stamp (Sawnat & Shah, 2013).

NoSQL Pattern

NoSQL databases can play a role with Hadoop implementation because NoSQL databases can store data on a local Network File System (NFS) disks as well as Hadoop Distributed File systems (HDFS). (Sawnat & Shah (2013). According to Network File System (2017), NFS is a distributed file system protocol originally developed by Sun Microsystems in 1984 allowing a user on a client computer to access files over a computer network much like local storage is accessed (Sawnat & Shah, 2013).

Polyglot Pattern

Allows multiple storage mechanisms such as RDBMS, Hadoop, and other Big Data appliances to co-exist in a solution. This scenario is known as "Polyglot Persistence" (Sawnat & Shah, 2013).

STORAGE MECHANISMS FOR BIG DATA

According to Chen et al. (2014), there are three bottom-up levels of storage mechanisms for Big Data: (1.) File Systems, (2.) Databases and (3.) Programming Models.

File Systems for Big Data

File systems are the foundations of the applications at upper levels. For example, Google File System (GFS) is an expandable distributed file system to support large-scale, distributed, data-intensive applications. (Chen et al., 2014). Facebook utilizes Haystack File System (HFS) to store the large amounts of small-sized photos. (Beaver et al., 2010). Table 2 below provides a list and description of some current File Systems for Big Data as derived from reading Chen et al. (2014) and Pierson (2017) the later of which claims the four of the best Big Data file systems are Hadoop Distributed File System (HDFS), Apache Spark, Quantcast and GlusterFS that are described below.

Database Technology for Big Data

There are three database technologies for Big Data that are to be discussed in this chapter and these are: 1.) Key-Valued Databases, 2.) Column-Oriented Databases and 3.) Programming Models for Big Data. Each of these are discussed briefly below and summarized in Table 3: Database Technology for Big Data Storage Systems.

Key-Valued Databases

One of the most recognized storage mechanisms is that used by Amazon for its processing of its huge data warehouse and is an example of a Key-Value Storage system and named Dynamo. Another Key-Value Storage system is named Voldemort and is used by LinkedIn. Other Key-Value Storage systems include Memcached, Riak, and Scalaris, Tokyo Canbinet and Tokyo Tyrant all of which provide expandability by dstribuiting key words into nodes (Chen et al., 2014).

Column-Oriented Databases

The column-oriented databases store and process data according to columns rather than rows. Columns and rows are segmented in multiple nodes to realize expandability. Examples include Google's BigTable, Cassandra and Derivatives of BigTable such as HBase and Hypertable (Chen et al., 2014).

Google's BigTable is a distributed, structured data storage system that is designed to process Big Data among thousands of commercial servers. (Chang et al., 2008) BigTable is based on many

Table 2. File systems for Big Data (Created by authors using Chen et al. (2014) and Pierson (2017) references.)

File Systems for Big Data	Characteristics
Azure Cosmos DB (Microsoft)	Microsoft's proprietary globally-distributed, multi-model database service launched in May 2017.
Facebook Haystack	Captive object storage system that in April 2009 managed 60 billion photos and 1.5 petabytes.
Google File System (GFS)	Distributed file system that uses Linux kernel operating system, released version named Colossus in 2010.
Taobao File Systems (TFS)	Used in China's biggest online shopping website similar to Amazon/EBay.
Hadoop Distributed File System (HDFS)	Use MapReduce as a key function of its data management. Product of Apache Software Foundation.
Apache Spark	Spark scalability is a key benefit.
QuantCast File System (QFS)	Specializes in measuring audience engagement on Internet sites, producing over 800,000 transactions per second over 100 million websites.
GlusterFS	Most notably used for cloud computing, steaming media, and content delivery.

fundamental components of Google such as Goggle File System (GFS), cluster management system, SSTable file format, and Chubby that among other functions also conducts error recovery in case of Table server failures (Chen et al., 2014).

Cassandra is another column-oriented database and was developed by Facebook and became an open-source tool in 2008. Tables in Casandra are in the form of distributed four-dimensional structured mapping, where the four dimensions are row, column family, column, and super column. Columns may constitute clusters that are called "column families" (Chen et al., 2014).

HBase and HyperTable are derivative tools of BigTable. HBase is a BigTable clone with java programming and is part of Hadoop pf Apache's MapReduce framework. The rows operations of HBase are operations with row-level locking and large-scope transaction processing (Chen et al., 2014).

Document Databases

Document storage databases can support more complex data forms than key-value storage. Examples of Document database storage systems are MongoDB, SimpleDB and CouchDB.

MongoDB is an open-source document-oriented database and supports horizontal expansion with automatic sharing to distribute data among thousands of nodes by automatically balancing load and keep the system up and running in case of failure (Chen et al., 2014).

SimpleDB is a distributed database and a web service of Amazon, but does not support automatic partition and thus cannot be expanded with the change of data volume, and also does not feature Multi-Version Concurrency Control (MVCC) that detects conflicts from other clients (Chen et al., 2014).

Platform for Nimble Universal Table Storage (PNUTS)

Platform for Nimble Universal Table Storage (PNUTS) is a large-scale parallel geographical-distributed system for Yahoo!'s web applications. In the physical layer of PNUTS, the system is divided into different regions each of which includes a set of complete system components and complete copies of tables. The data table is horizontally segmented into record groups which are called Tablets. Tablets are distributed among many servers each of which may have tens of thousands of Tablets, but a Tablet may only be stored in a region of a server. (Chen et al., 2014).

Programming Models for Big Data

According to Chen et al. (2014), the traditional parallel models of Message Passing Interface (MPI) and Open Multi-Processing (OpenMP) may not be adequate to support large-scale parallel systems with

Table 3. Database Technology for Big Data Storage Systems (Created by authors using Chen et al. (2014) reference)

Key-Value Databases	Column-Oriented Databases	Document Databases	Platform-Based Databases
Dynamo	Big Table [derives HBase & HyperTable]	MongoDB	Platform for Nimble Universal Table Storage (PNUTS)
Voldemort	HBase	SimpleDB	
Tokyo Canbinet	HyperTable	CouchDB	
Tokyo Tyrant	Cassandra		

hundreds and thousands of massive datasets of Big Data that are stored in hundreds or thousands of commercial servers. Hence there is the need for Big Data Programming models such as MapReduce, Dryad, All-Pairs and Pregel.

MapReduce

MapReduce is a simple but powerful programming model for large-scale computing using a large number of clusters of commercial PC's to achieve automatic parallel processing and distribution. In MapReduce, the computational workloads are performed by inputting key-value pair sets and generating key-value pair sets. The computing model has only two functions both of which are programmed by users: 1.) Map 2.) Reduce. The Map function processes input and generates intermediate key-value pairs. Then MapReduce will combine all the intermediate values related to the same key and transmit them to the Reduce function, Next the Reduce function receives the intermediate key and its value set, merges them, and generates a smaller value set (Chen et al., 2014).

Dryad

Dryad is a general-purpose distributed execution engine for processing parallel applications of coarse-grained data. The operational structure of Dryad is a directed acyclic graph, in which vertexes represent programs and edges represent data channels. Dyad executes operations on the vertexes in computer clusters and transmits data via data channels, including documents, TCP connections, and shared-memory FIFO (First In-First Out) (Chen et al., 2014).

All-Pairs

All-Pairs is a system specifically designed for biometrics, bio-informatics and data mining applications, All-Pairs focuses on comparing element pairs in two databases by a given function. The All-Pairs problem maybe expressed as a three-tuples (Set A, Set B, and Function F), in which Function F is utilized to compare all elements in Set A and Set B. The comparison results in an output matrix M called the Cartesian product or cross-join of Set A and Set B (Chen et. al., 2014).

Pregel

The Pregel system of Google facilitates the processing of large-scale graphs. Applications include that for analysis of network graphs and social networking services. The Pregel program output is a set consisting of the values output from all the vertexes, and the Pregel program output and input are an isomorphic directed graph. (Chen et. al., 2014). Table 4 shows programming models for big data.

Wu et al. (2017b) discusses a taxonomy of Big Data programming models some of which is partially presented below in Table 5.

Table 4. Programming models for Big Data (Created by authors using Chen et al. (2014) reference)

Big data programming models	Characteristics	Applications
All-Pairs	Cartesian product of (Set A, Set B, Function F)	Biometrics, Bioinformatics
Dryod	Directed Acyclic graph	Parallel-processing of coarse-grained data.
MapReduce	Computing model uses only functions: Map, Reduce	Genomics analysis
Pregel	Google based that facilities the processing of large-scale graphs.	Network Graphs

TOP 10 IN THE WORLD CLOUD STORAGE SYSTEMS FOR DATA-INTENSIVE COMPUTING USING BIG DATA

According to a Gartner survey, about 19% of organizations are using the cloud for production computing, while 20% are using public cloud storage services. Gartner surveyed 556 organizations, from June 2012 through July 2012, across nine countries and multiple industries where cloud planning is a critical issue.

The survey of Gartner (2012) found that public cloud adoption varies by service. IaaS is moving from lower-risk pilot programs and into production environments. Organizations' stated plans to adopt IaaS in the near future reinforce the importance of IaaS in an overall portfolio of infrastructure service offerings.

According to Butler (2013), the following Table 6 lists The Top 10 Cloud Storage Providers as having been ranked by a Gartner Report.

Below are some brief characteristics of these Top 10 Storage Providers as obtained from Bulter (2013) to which the reader is referred to for more complete information.

Amazon Web Services (AWS)

Like many other aspects of cloud computing, Amazon Web Services is considered a market leader in cloud storage. Gartner says, while its pricing is the "industry reference point." Its Simple Storage Service (S3) is the basic object storage, while Elastic Block Storage is for storage volumes. AWS has introduced RedShift, a cloud-based data warehousing service, and AWS Storage Gateway that has the ability to create hybrid storage architectures that span both on-premise storage options and AWS's cloud that is still largely a work in progress. (Butler, 2013).

AT&T

AT&T's Synaptic cloud storage service is aligned closely with EMC's Atmos storage service, which is used as an on-premise storage system. This creates an opportunity for AT&T to sell into the strong

Table 5. Taxonomy of Programming Models (Derived from Wu et al. (2017b)

MapReduce	Functional	SQL-based	High-level DSL
MapReduce	Spark	HiveQL	Pig Latin
Hadoop	Flink	CasandraQL	Cascading
		SparkSQL	LINQ
		Impala	Trident

Table 6. Top 10 Cloud Storage Providers (Source: Butler (2013).)

STORAGE RANK	STORAGE SYSTEM	HEADQUARTERS
1	Amazon Web Services	Seattle, WA USA
2	AT & T Synaptic Cloud Storage	Dallas, TX USA
3	Google Cloud Storage	Mountain View, CA USA
4	Hewlett-Packard (HP) Cloud Open Stack	Palo Alto, CA USA
5	IBM Smart Cloud Enterprise	Armonk, NY USA
6	Internap AgileFiles	Atlanta, GA USA
7	Microsoft Windows Azure Blob	Redmond, WA USA
8	Nirvaniz	San Diego, CA USA
9	Rackspace Cloud Block Storage	Windcrest, TX USA
10	Softlayer Cloudlayer	Dallas, TX USA

EMC customer base, and gives customers hybrid cloud capabilities with a leading storage vendor. (Butler, 2013).

Google Cloud Storage

Launched in 2010, Google Cloud Storage is the underlying storage service for the company's other cloud products and services, including Google App Engine - the application development platform - Google Compute Engine, and BigQuery, which are cloud-based virtual machines and a Big Data analysis tool, respectively. (Butler, 2013).

Hewlett-Packard (HP)

"Among OpenStack-based cloud storage providers, HP is well-positioned to understand enterprise IT storage requirements, due to its extensive hardware, software and service offerings," Gartner notes. "However, since HP Cloud Object Storage is new, HP must evolve and refine its architectural, geographical and service offerings." The system automatically replicates data across three availability zones for resiliency (which customers can choose to do in Amazon's cloud), and HP says having information running on its hardware both in the public cloud and on customers' premises makes for easy hybrid cloud setups (Butler, 2013).

IBM

IBM's cloud storage is part of its SmartCloud Enterprise offering, which includes other services such as cloud-based application development and infrastructure. Gartner says the biggest deficiency is the lack of integration among the various aspects of IBM's SmartCloud offering though. For example, IBM markets its cloud for backup and recovery, but those services do not use IBM SmartCloud Object Storage on its backend. Part of this could be because IBM partners with Nirvanix, another cloud storage provider, to run the SamrtCloud Object Storage. (Butler, 2013).

Internap

To differentiate its service, Internap has attempted to layer on advanced networking features to the service, such as a Manager Internet Route Optimizer (MIRO), which analyzes the performance of the possible routes to deliver content and chooses the best one. Gartner says its lack of enterprise presence is the biggest limitation holding the company back (Butler, 2013).

Microsoft

Behind Amazon Web Services, Microsoft's Windows Azure Blob Storage may be the second most widely-used cloud storage service, Gartner predicts. It now boasts more than a trillion objects and is growing at 200% per year, Gartner says, while supporting a broad range of features including object storage, table storage, SQL Server and a content delivery network (CDN). (Butler, 2013).

Nirvanix

Nirvanix has some appealing characteristics though, including the ability to have public, hybrid or on-premise Nirvanix-powered storage services, and an all-inclusive monthly billing cycle with premium support options, a clear aim for enterprise customers, but one that may turn away small and midsized businesses that may prefer the a la carte pricing (Butler, 2013).

Rackspace

For high-performance storage needs, it has Cloud Block Storage, which has high input-output capabilities. Rackspace works heavily on the OpenStack open source project and its services closely follow OpenStack developments. Because of its work in the OpenStack environment, Gartner says Rackspace public cloud storage services integrate nicely with OpenStack-powered-on premise clouds, creating hybrid cloud services for customers (Butler, 2013).

Softlayer

Softlayer also has a storage-area network (SAN) offering and an international presence, with data center locations in its headquarters of Dallas, along with Amsterdam and Singapore. Its lack of support and turnkey deployment cycles, Gartner says, has meant that the product has not caught on wildly with the enterprise market yet though (Butler, 2013)

TOP TEN BIG DATA STORAGE TOOLS

There are many Big Data Storage Tools on the market and there is no simple answer as which ones are the best. However, Robb (2016) created a list of the "Top Ten Big Data Storage Tools" from which the below Table 7 of important characteristics was derived by the authors of this chapter upon reading this article.

The important characteristics of the above "Top Ten Big Data Storage Tools" range from supporting Big Video by Hitachi to detecting data errors in real-time to optimize the performance of Big Data projects using Infogix.

There are many variables in selecting a Big Data Storage tool, and these include according to Robb (2016) the existing environment, current storage platform, growth expectations, size and type of files, database and application mix, among other variables specific the unique needs of the users.

BIG DATA STORAGE FOR TOP 500 SUPERCOMPUTERS IN THE WORLD

Big Data storage is essential for the architecture of any supercomputer. The Top500 Supercomputers in the World were discussed in detail in earlier IGI Global book by Segall, Cook and Zhang (2015) and with Chapter 1 by Segall and Gupta (2015) titled *Overview of Global Supercomputing,* and summarized statistics in Appendix: The Top 500 Supercomputers in the World by Gupta (2015).

The TOP500 project that was started in 1993 ranks and details the 500 most powerful non-distributed computer systems in the world. In a recent list (June 2017), the Chinese Sunway TaihuLight is the world's most powerful supercomputer, reaching 93.015 petaFLOPS on the LINPACK benchmarks. (Top 500, 2018a)

Table 7. Characteristics of Top Ten Big Data Storage Tools (Created by authors with reference to Robb (2016).)

DATA STORAGE TOOL	IMPORTANT CHARACTERISTICS
1. Hitachi	Hitachi Video Management Platform (VMP) supports Big Video.
2. DDN (DataDirect Networks)	High performance file storage can be automatically tiered to object storage archive to support cost-effective retention of Big Data.
3. Spectra BlackPearl	Object storage interface to SAS-based disk
4. Kaminardo K2	All-flash array to support dynamic workload
5. Caringo	Flagship product Swarm eliminates the need to migrate data into disparate solutions for long-term preservation, delivery and analysis thereby lowering total cost of ownership.
6. Infogix	Can detect data errors in real-time to optimize the performance of Big Data projects.
7. Avere Hybrid Cloud	Users can harness object storage without rewriting their applications or changing their data access methods.
8. DriveScale	DriveScale allows users to procure capacity independent of the compute capacity thus enabling right-sizing at each level.
9. Hedvig	Users can customize storage with a range of enterprise services that are selectable per-volume.
10. Nimble	Nimble Storage Predictive Flash Platform dramatically improve performance of analytical applications and Big Data workloads.

The TOP500 list is compiled by Jack Dongarra of the University of Tennessee, Knoxville, Erich Strohmaier and Horst Simon of the National Energy Research Scientific Computing Center (NERSC) and Lawrence Berkeley National Laboratory (LBNL), and also by Hans Meuer of the University of Mannheim, Germany until his untimely death in 2014. (Top 500, 2018a)

This section provides some insight into Big Data storage by providing some of the statistics of storage for Big Data for the Top500 Supercomputers. The reader is referred to the web-site https://www.top500.org/statistics/list/ that allows the user to drill-down for additional statistics and visualization plot of storage related capabilities of components of the Top 500 Supercomputers in the World in 2018.

Below are graphical illustrations and explanations for June 2018 statistics pertaining to Big Data storage on the Top 500 Supercomputers in the World.

Cores per Socket

Table 8 is a tabular summary that indicates that 16 Cores per Socket is the most frequent System Share with 21.8% of the Top500 Supercomputers that have 9,568,908 cores and the least frequent is 260 cores with .2% system share.

Figure 2 consists of two pie charts: One that represents the Cores per Socket Share and the other Cores per Socket Performance Share. From these two pie charts we can visualize the greatest Core per Socket System Share is 16 or 21.8% and the lowest percentage is for 6 Cores per Socket System Share, and the greatest Cores per Socket Performance is 16 or 26.2% and the lowest is for 6 cores per socket. Figure 3 shows the TreeMap of cores per socket as of June 2018.

TreeMaps display hierarchical (tree-structured) data as a set of nested rectangles. Each branch of the tree is given a rectangle, which is then tiled with smaller rectangles representing sub-branches. A leaf node's rectangle has an area proportional to a specified dimension on the data.

Figure 4 shows that the majority of Efficiency (%) were at least 50% for all of the number of cores per socket in 2018 of the Top 500 supercomputers in the world.

Table 8. Top 500 Big Data storage servers list statistics by Cores per socket table as of June 2018. (Top 500, 2018b) (Source: https://www.top500.org/statistics/list/)

Cores per Socket	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
8	34	6.8	46,781,736	72,213,114	4,901,860
68	11	2.2	93,474,699	192,235,629	4,229,540
64	8	1.6	17,248,140	30,219,980	726,960
6	11	2.2	13,477,800	26,760,586	776,548
32	5	1	10,422,200	11,530,310	354,384
260	1	0.2	93,014,594	125,435,904	10,649,600
24	15	3	55,781,560	89,381,831	1,229,744
22	5	1	196,897,384	311,144,821	3,954,192
20	69	13.8	127,487,637	183,701,065	3,620,672
18	43	8.6	83,997,092	104,750,192	2,802,792
16	109	21.8	167,676,581	270,106,377	9,568,908
14	41	8.2	51,410,355	85,228,710	1,863,500
12	105	21	202,000,027	323,987,492	10,884,042
10	43	8.6	51,245,059	94,968,791	2,492,892

Vendors Systems & Performance Share

Figure 5 indicates that Lenovo has the largest Vendor System Share of 23.4% and IBM, Huawei, Dell, Fujitsu and others have the least Vendor System Share for the Top 500 supercomputers of 4% or less.

Figure 6 shows that the Vendors System Share for IBM and Hitachi have been declined significantly from 2006 to 2018 for the Top 500 Supercomputers in the world, while that of Cray, Bull, Dell, HPE and others have both increased and decreased over this time period.

Figure 7 indicates that Lenovo has the largest Vendor Performance Share of 19.9% followed close by HPE of 19.7%, and Huawei, Dell, Fujitsu and others have the least of less than 5%.

Figure 8 shows that the Vendor Performance Share for HPE had the most of the Share (%) over the time period of 2006-2018, and IBM had a declining share over this period, and was below other vendors such as Oracle.

Table 9 provides a table of vendors that show that Lenovo, HPE and Inspur, Sugon, and Cray, Inc. combined have 74.4% of the System Share of the Top500 supercomputers while 16 of the listed vendors have .2%, and 8 of the listed vendors have .4%. Figure 9 shows the TreeMap of vendors Rmax as of June 2018.

Rmax and Rpeak values are in GFlops. Rpeak values are calculated using the advertised clock rate of the CPU. For the efficiency of the systems you should take into account the Turbo CPU clock rate where it applies. Cray Inc. has the largest and HPE the second largest share followed by NRCPC, Levovo and IBM.

Application Area & System Share

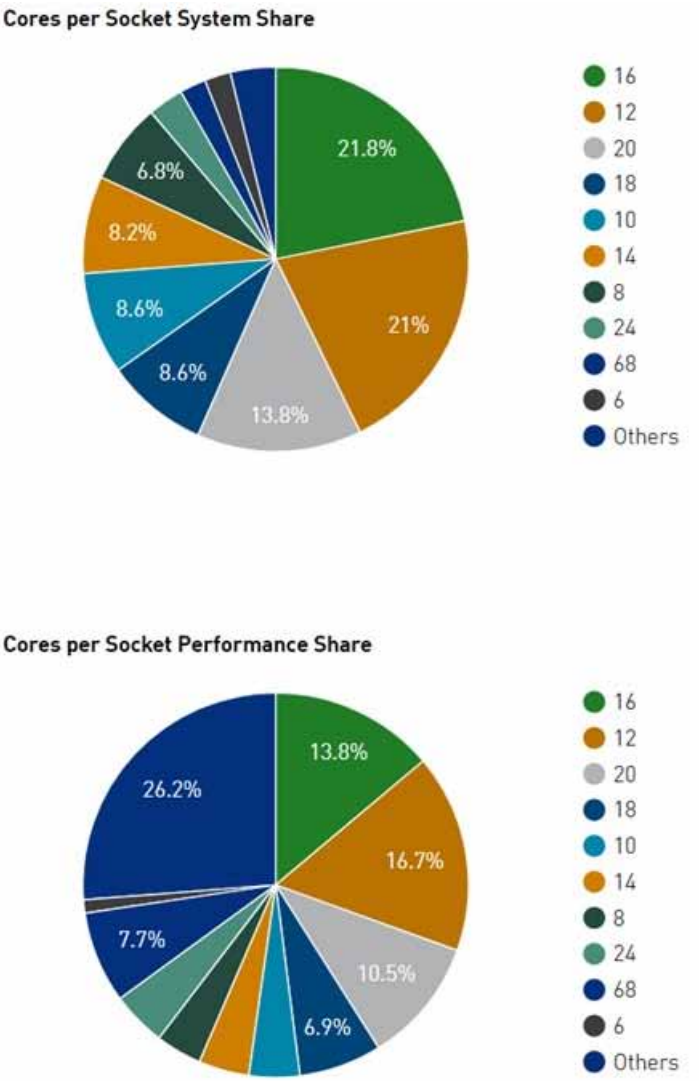
Figure 10 indicates that 96.6% or essentially all of the *Application Area System Share* for *the Top 500 Supercomputers is unspecified, and 89.6% of the Application Area Performance Share* is also unspecified.

Table 10 indicates that 483 of the Top 500 Supercomputers in the world have an *Application Area* that is Not Specified, and these are 96.6% of the System Share and consist of 45,218,714 Cores.

Architecture

Figure 11 indicates that 87.4% of the *Architecture System Shar*e is Cluster and 12.6% is MPP (Massive Parallel Processing), and 74.4% of the *Architecture Performance Share* is Cluster and 25.6% is MPP.

Figure 2. Top 500 Big Data storage servers pie-chart statistics by Cores per socket as of June 2018. (Top 500, 2018b) (Source: https://www.top500.org/statistics/list/)



Operating System

Figure 12 indicates that for the Top 500 Supercomputers in the world in 2018, Linux had the majority of *Operating Systems System Share* with 50.8% and CentOS with 23.2% and Cray Linus Environment with 9.8%, and *Operating System Performance Share* was dominated by Linux with 29.5%, CentOS with 18% and Cray Linux Environment with 12.5%.

Table 12 provides statistics of numerical counts of the types of Operating Systems for the Top 500 Supercomputers in the world in 2018 with a count of the number of cores. From Table 12 it can be noted that the one supercomputer in the world with .2% of System Share of Sunway RaiseOS 2.0.5 is located at National Supercomputer Center in China at Wuxi. Jiangsu has the second largest number of cores of 10,649,600 as compared to the 254 that has system share of 50.8% of Supercomputers with Linux Operating System that has the maximum number of cores of 18,422,444.

Figure 3. TreeMap of Cores per Socket as of June 2018 (Top 500, 2018c)

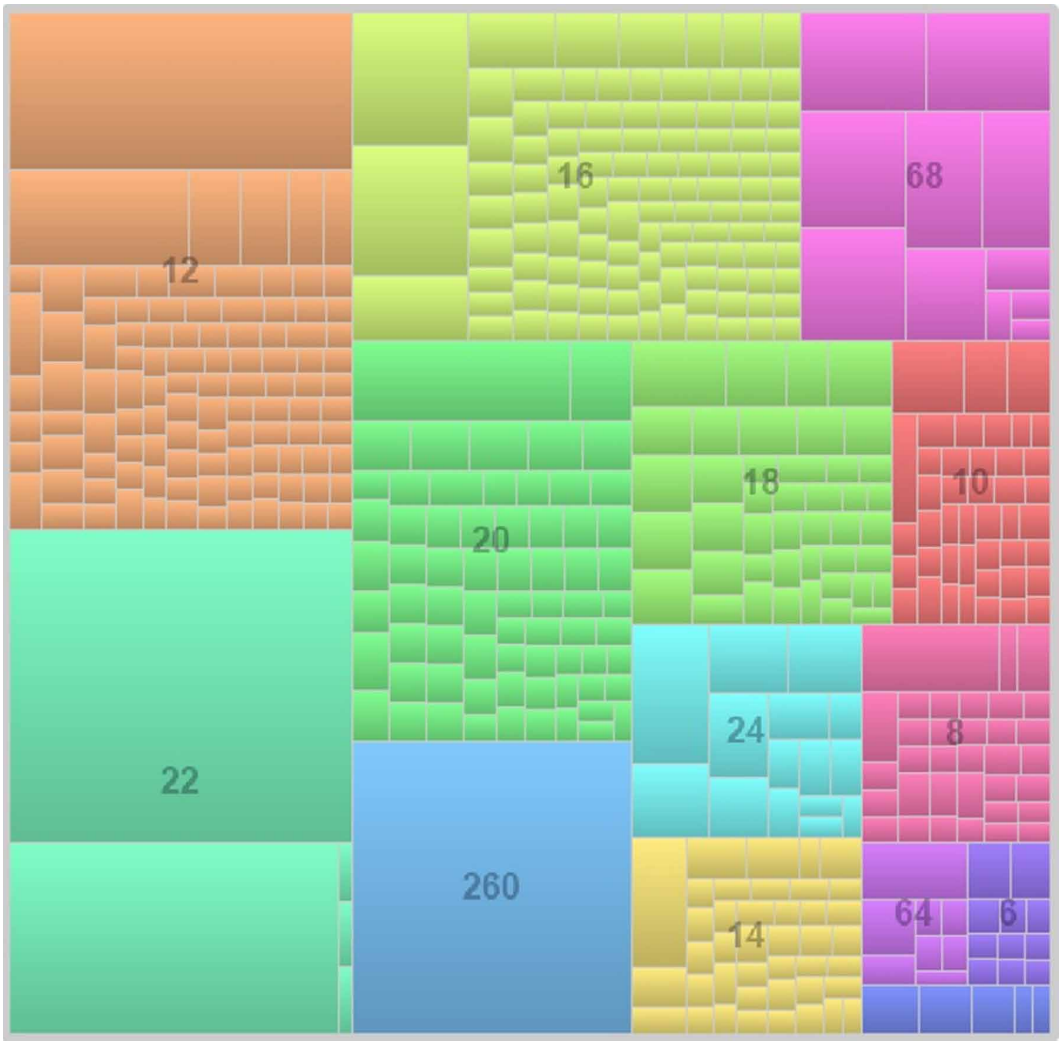


Figure 13 shows that all of the *Operating System Family Systems Share* and *Operating System Family Performance Share* of the Top 500 Supercomputers in the World in 2018 were that of Linux. This was an increase from the June 2017 data when for System Share was 99.6% for Linux and the remaining .04% were Unix., and 99.9% Performance Share was Linux and the remaining .1% were other.

Table 13 indicates that the Linux Operating Systems Family that that comprised 99.6% of the Top500 Supercomputers in 2018 had 48,043,950 Cores of storage.

Accelerator/Co-Processor

Figure 14 indicates that NVIDIA Telsa K40, NVIDIA Telsa P100, NVIDIA Telsa K80 and NVIDIA Telsa 20x comprised about 73% of the Accelerator/Co-Processor System Share.

Table 14 indicates that NVIDIA Telsa K40 had the maximum system share, but Intel Xeon Phi 31S1P supercomputer located at National Super Computer Center in Guangzhou, China had the maximum number of storage cores of 3, 294,720 followed by PEZY-SC2 that had 3.176,00 cores of storage and processing and the PEZY-SCnp is said to deliver 1.53 TFLOPS (double-precision).

Figure 4. Efficiency chart of cores per socket as of June 2018 (Top 500, 2018d) (Source: https://www.top500.org/statistics/efficiency-power-cores/)

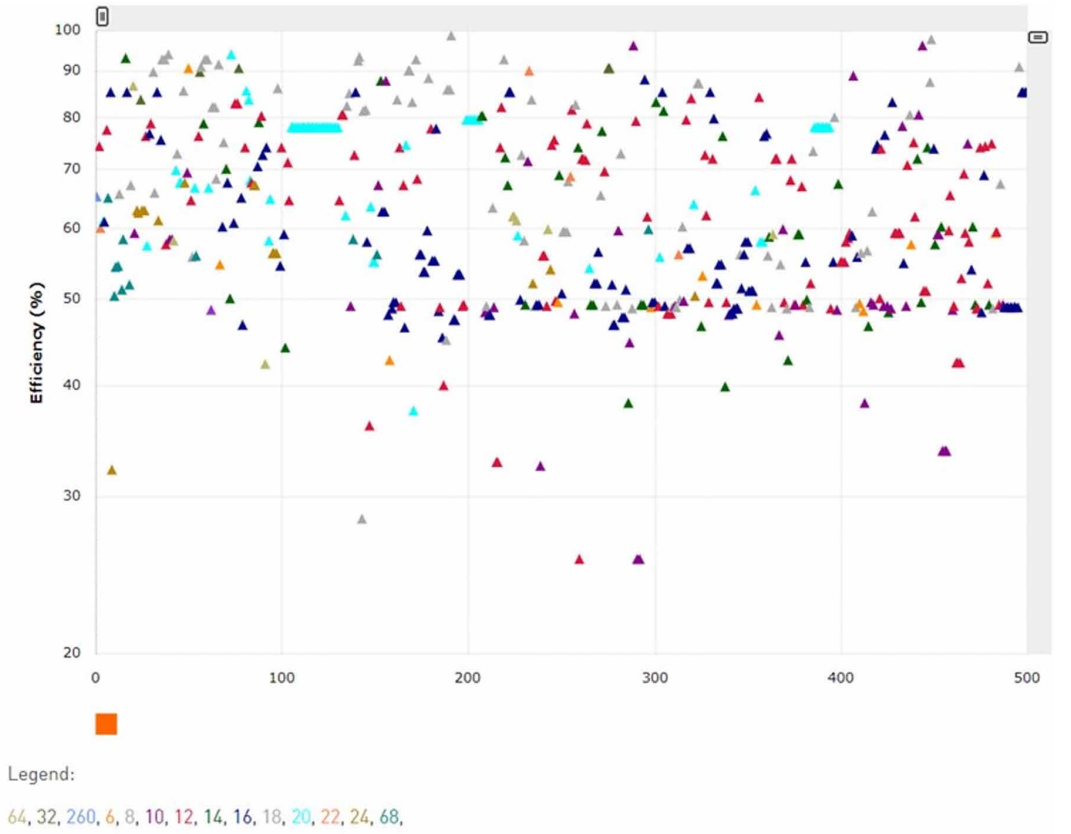


Figure 5. Top 500 Big Data storage servers list statistics by Vendors Systems Share as of June 2018. (Top 500, 2018b) (Source: https://www.top500.org/statistics/list/)

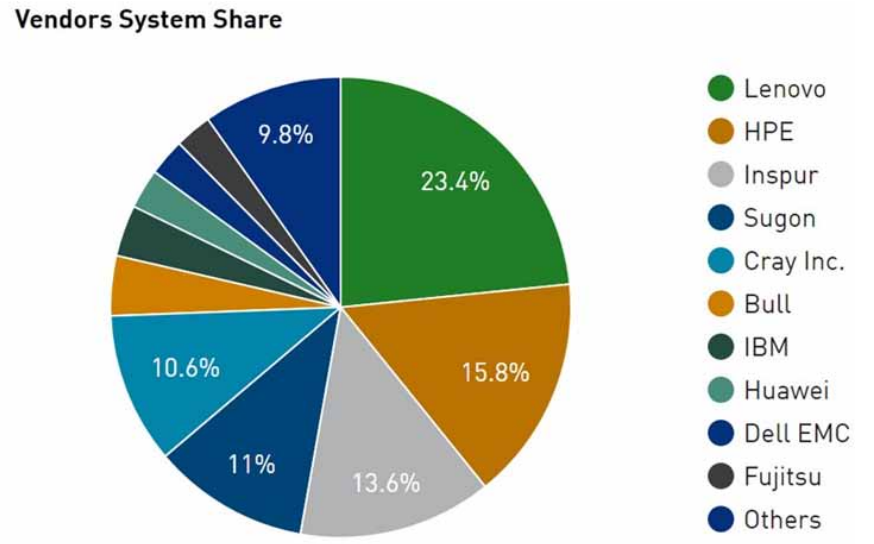
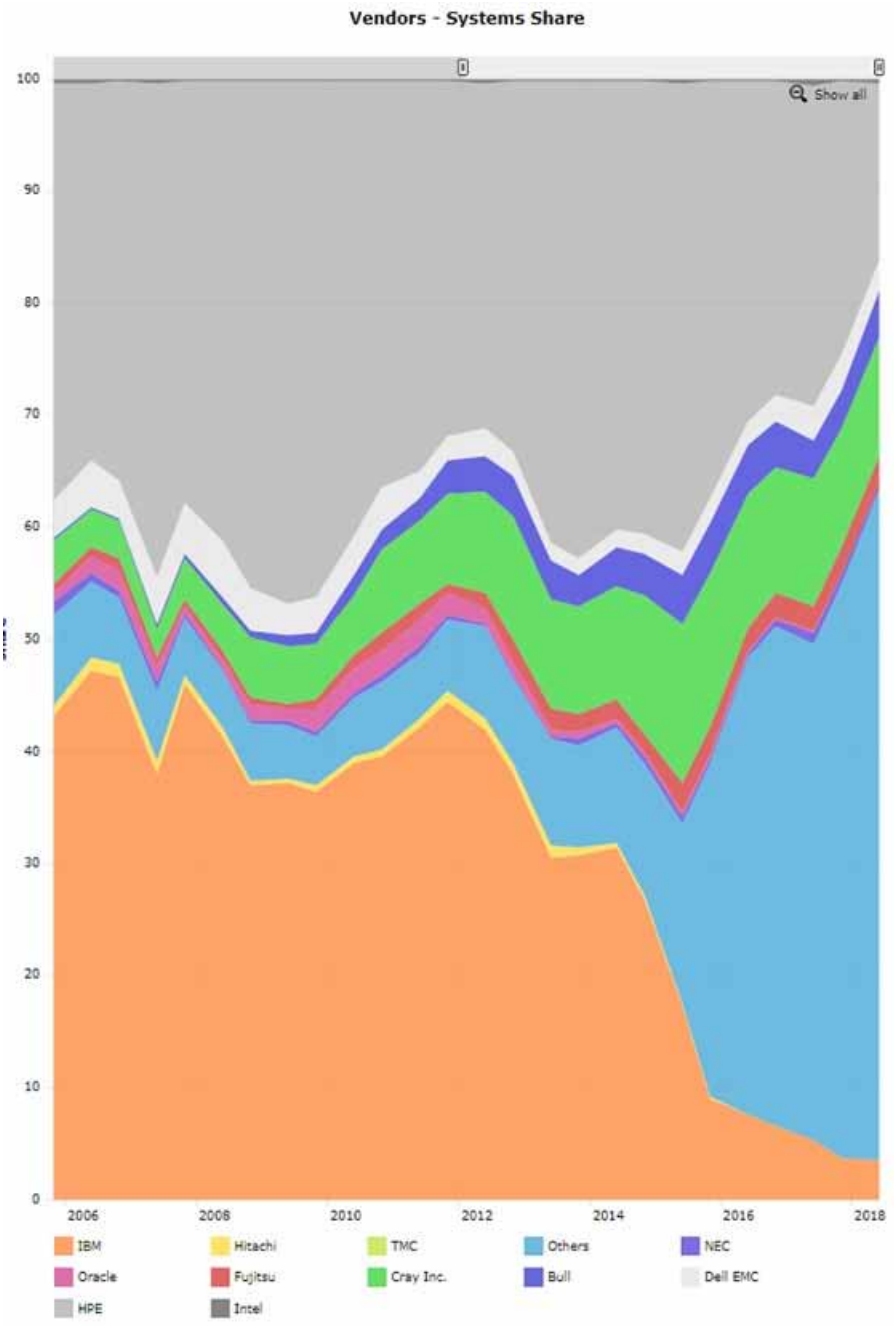


Figure 6. Development over time by Vendors Systems Share as of June 2018. (Top 500, 2018e) (Source: https://www.top500.org/statistics/overtime/)



Interconnect System

Figure 15 indicates that the majority of the *Interconnect System Share* of the Top 500 supercomputers was either 10G Ethernet or Infiniband FDR and the *Interconnent Performance Share* was dominated by 10G Ethernet, Infinibanc FDR, and Intel Omni-Path.

Figure 7. Top 500 Big Data storage servers list statistics by Vendors-Performance share as of June 2018 (Top 500, 2018b) (Source: https://www.top500.org/statistics/list/)

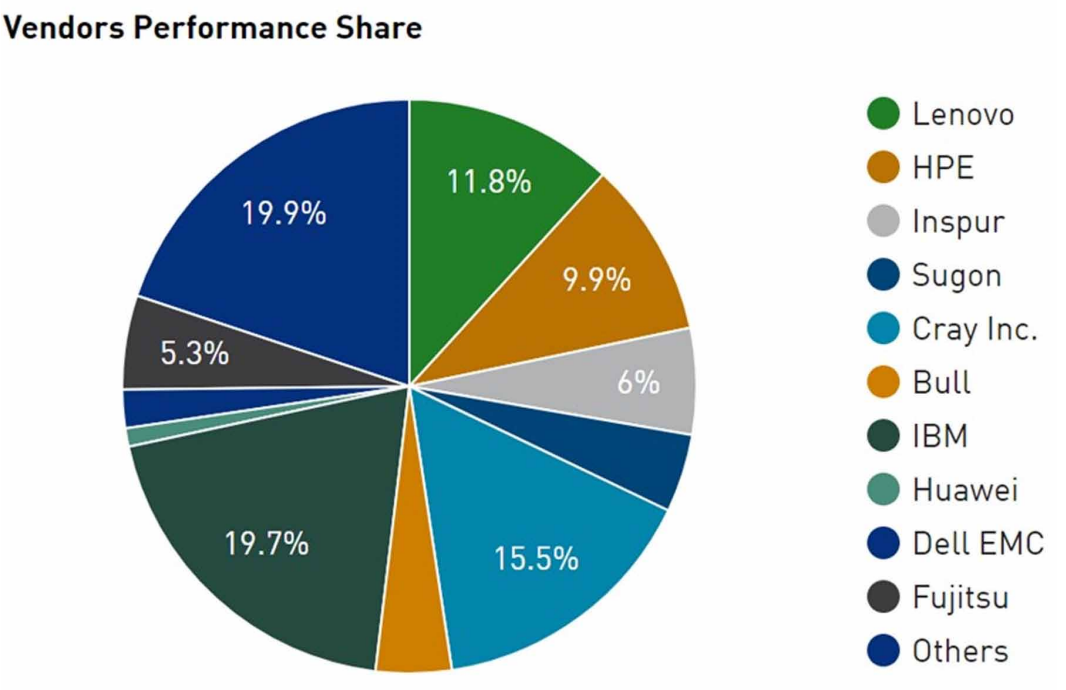


Table 15 indicates that the Sunway microprocessors developed in China had the maximum number of processing and storage cores of 10,649,600.

CONCLUSION

The topic of storage systems for data-intensive computing of Big Data that utilizes Cloud and Fog Computing is a topic of crucial importance to an increasingly very many number of organizations around the world, be it public, private, government, or industrial.

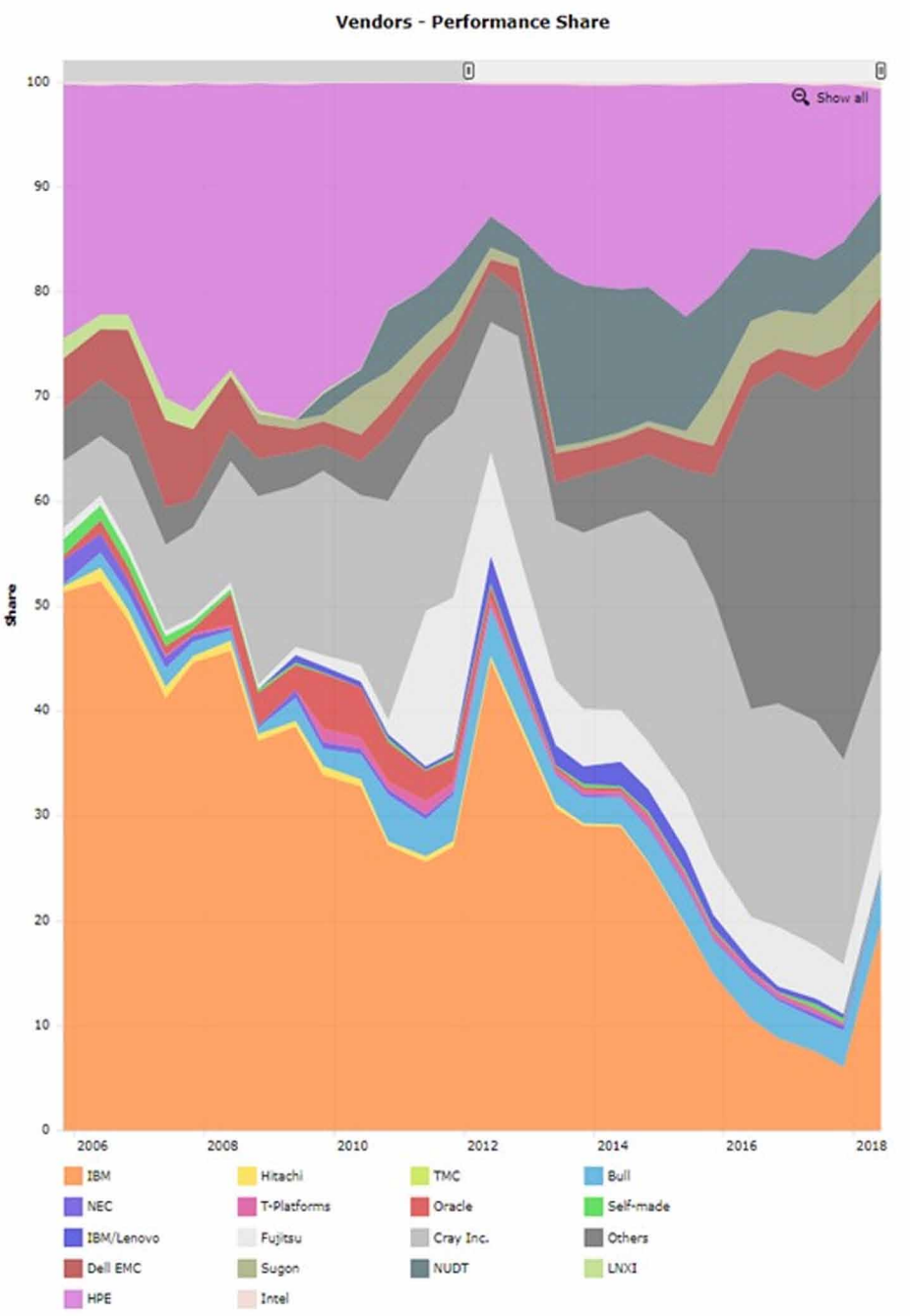
The topic and importance of "real-time" processing of Big Data also is an issue of great importance and is also discussed in Sakr (2016) and Swami et al. (2018). Figure 16 illustrates "*real-time"* Big Data Processing with input/output centric storage with a five-layer model for future computing for parallel processing consisting of transactional, analytic, operational and archive systems. Belli et al. (2018) studied scalable Big Stream Cloud Architecture for IoT.

Figure 17 below identifies 36 specific areas that are impacted by data growth as a "ripple effect". As Cochran (2012) indicates the importance of Big Data storage and processing by stating:

"In today's world of 'Big Data', there needs to be far greater emphasis on comprehensive planning, designing in architectural efficiency, minimizing the impact on IT infrastructure, and improving the manageability of our entire IT environment. Your future depends on it."

In conclusion, storage systems for data-intensive computing using Big Data for Fog Computing and its applications using Fog-to-Cloud Computing is an expanding and vital topic that needs to be pursued continuously for the financial successes of economies around the world.

Figure 8. Development over time by Vendors-Performance Share as of June 2018 (Top 500, 2018e) (Source: https://www.top500.org/statistics/overtime/)



FUTURE DIRECTIONS OF RESEARCH

The future directions of this research include the continuation of pursuing knowledge of updates in the technology for storage systems for data-intensive computing using Big Data and its interface

Table 9. Top 500 Big Data storage servers list statistics by Vendors table as of June 2018. (Top 500, 2018b) (Source: https://www.top500.org/statistics/list/)

Vendors	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
Lenovo	117	23.4	142,630,714	244,372,439	5,997,344
HPE	79	15.8	120,436,184	175,633,682	4,116,904
Inspur	68	13.6	72,503,600	141,667,027	2,472,544
Sugon	55	11	53,316,500	102,952,928	2,840,928
Cray Inc.	53	10.6	187,798,219	297,606,643	7,158,496
Bull	21	4.2	51,623,508	80,264,489	1,913,640
IBM	18	3.6	239,067,376	360,999,821	7,543,968
Huawei	14	2.8	12,446,617	18,993,139	473,536
Dell EMC	13	2.6	26,238,810	43,597,252	813,508
Fujitsu	13	2.6	64,056,640	96,412,967	2,206,264
Penguin Computing	11	2.2	15,370,527	17,730,300	537,840
NUDT	4	0.8	66,853,590	109,796,949	5,396,096
PEZY Computing /	4	0.8	3,481,327	4,871,393	3,664,704
Intel	2	0.4	7,075,800	12,300,800	173,200
Cray Inc./Hitachi	2	0.4	11,461,000	18,250,444	271,584
Dell EMC / IBM-GBS	2	0.4	2,242,260	2,795,520	67,200
NEC	2	0.4	2,746,620	5,221,568	61,952
T-Platforms	2	0.4	3,379,900	6,647,000	143,044
Lenovo/IBM	2	0.4	4,096,932	5,041,434	151,336
Nvidia	2	0.4	4,377,000	6,716,264	82,952
Self-made	1	0.2	3,307,000	4,896,512	60,512
T-Platforms, Intel, Dell	1	0.2	3,782,570	6,563,840	155,150
SuperMicro/Mellanox	1	0.2	755,500	1,268,363	22,700
NSSOL/HPE	1	0.2	770,400	1,243,776	16,320
NRCPC	1	0.2	93,014,594	125,435,904	10,649,600
NRCPC	1	0.2	795,900	1,070,160	137,200
MEGWARE	1	0.2	800,327	1,277,107	17,352
Megatel/Action	1	0.2	1,010,940	1,413,120	38,400
Hitachi/Fujitsu	1	0.2	1,018,000	1,502,236	222,072
Cray Inc./T-Platforms	1	0.2	1,200,350	1,293,005	35,136
NTT Comm. / NTT PC	1	0.2	1,391,000	4,917,658	59,392
Fujitsu / Lenovo / Xenon	1	0.2	1,676,220	3,801,424	87,224
NEC/MEGWARE	1	0.2	1,967,810	2,800,870	49,432
Atipa	1	0.2	2,539,130	3,388,032	194,616
NEC/HPE	1	0.2	2,785,000	5,735,685	76,032
IBM/Lenovo	1	0.2	2,897,000	3,185,050	147,456

with Fog Computing. The constant generation of discoveries by investigators and authors around the world in this topic is of vital importance to the economic growth and well-being of all either directly or indirectly.

ACKNOWLEDGEMENT

This is an Invited Paper by Editors of the *International Journal of Fog Computing* (IJFC) published by IGI Global as an updated/expanded version of an earlier publication by the authors. An earlier chapter authored by R.S. Segall and J.S. Cook was titled "Overview of Big Data-Intensive Storage and its Technologies" that appeared as Chapter 2 in 2-volume Handbook of Big Data Storage and Visualization Techniques edited by R.S. Segall and J.S. Cook, and published by IGI Global in 2018 (ISBN-13: 978-1-522-53142-5 and e-Book EISBN-13: 978-1-522-53143-2).

Figure 9. TreeMap of Vendors Rmax as of June 2018 (Top 500, 2018c)

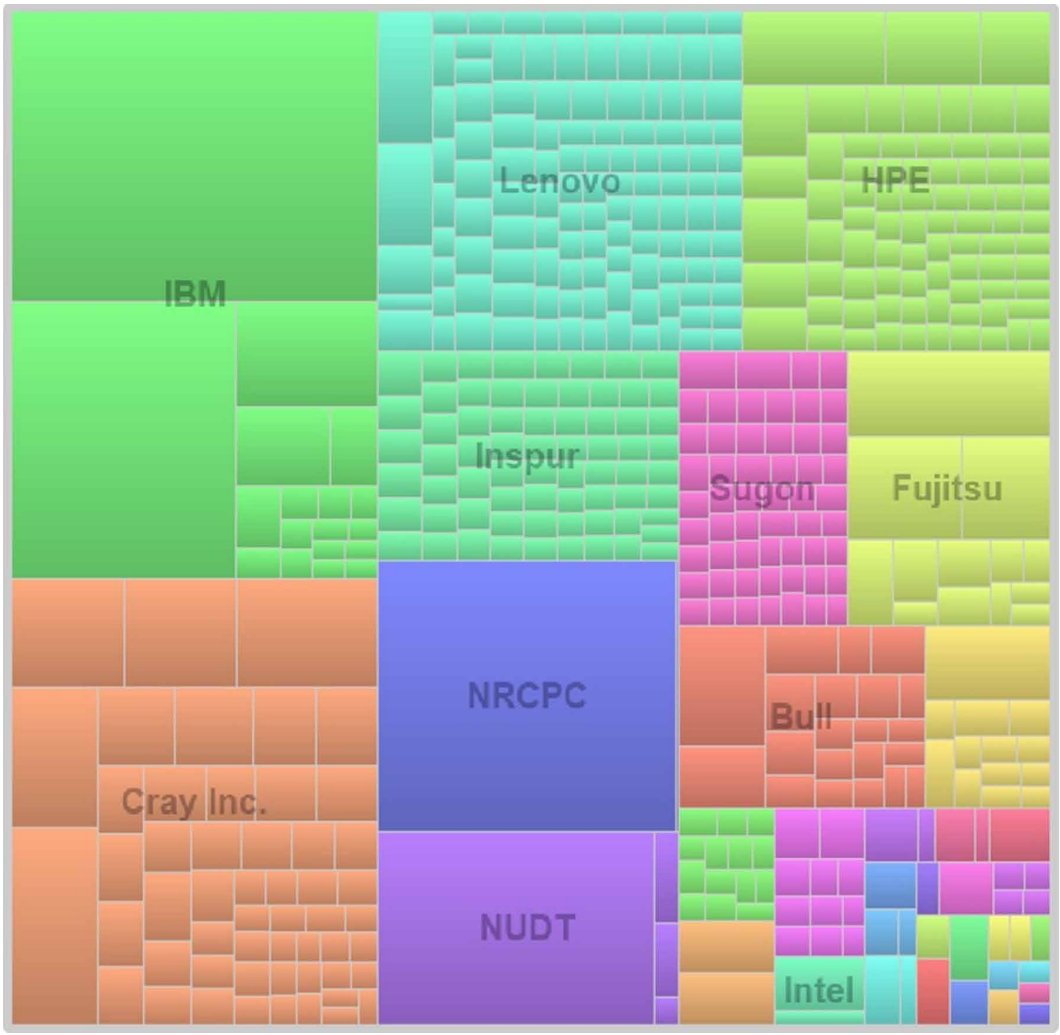


Figure 10. Top 500 Big Data storage servers list statistics by Application Area as of June 2018. (Top 500, 2018b) (Source: https://www.top500.org/statistics/list/)

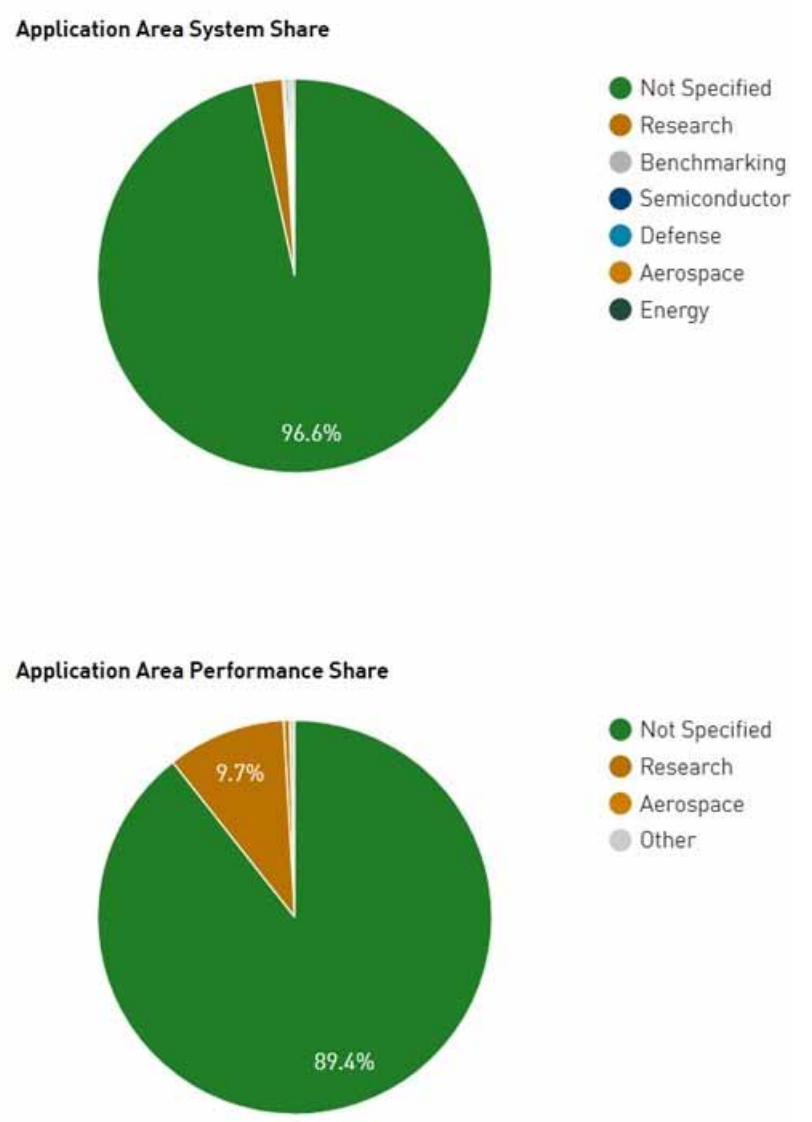


Table 10. Top 500 Big Data storage servers list statistics by Application Area table as of June 2018. (Top 500, 2018b). (Source: https://www.top500.org/statistics/list/)

Application Area	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
Not Specified	483	96.6	1,082,320,079	1,750,109,894	45,218,714
Research	12	2.4	117,600,436	157,618,909	12,319,252
Benchmarking	1	0.2	1,587,000	1,931,625	62,944
Semiconductor	1	0.2	1,462,970	2,507,264	45,680
Defense	1	0.2	1,050,000	1,254,550	138,368
Aerospace	1	0.2	5,951,550	7,107,149	241,108
Energy	1	0.2	942,829	1,135,411	29,568

Figure 11. Top 500 Big Data storage servers list statistics by Architecture as of June 2018. (Top 500, 2018b). (Source: https://www.top500.org/statistics/list/)

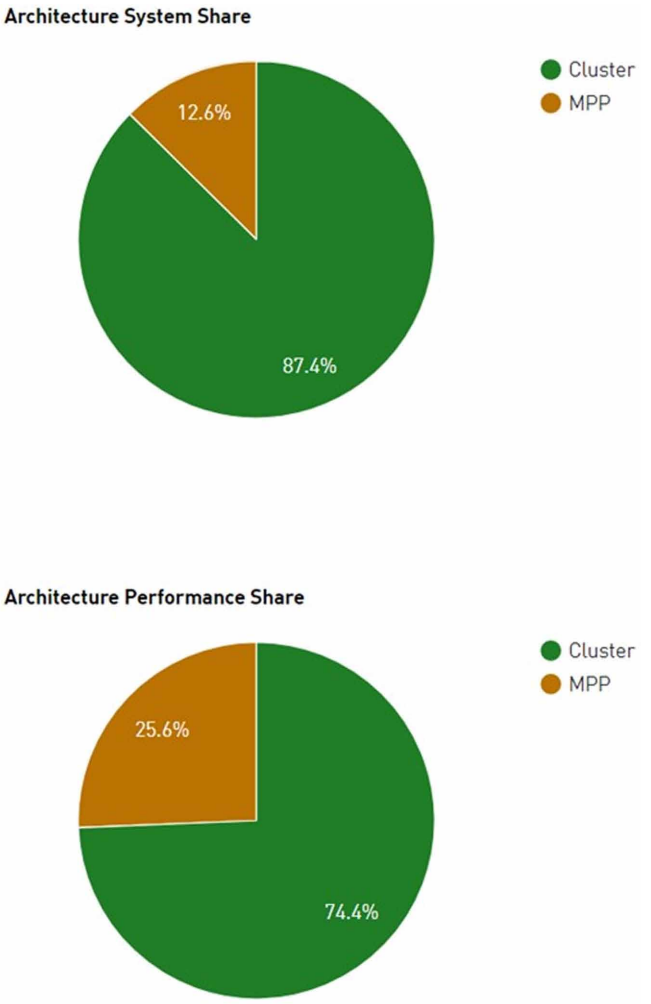


Table 11. Table of Top 500 Big Data storage servers list statistics by Architecture as of June 2018. (Top 500, 2018b). (Source: https://www.top500.org/statistics/list/)

Operating system Family	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
Linux	500	100	1,210,914,864	1,921,664,802	58,055,634

Figure 12. Top 500 Big Data storage servers list statistics by Operating System as of June 2018 (Top 500, 2018b). (Source: https://www.top500.org/statistics/list/)

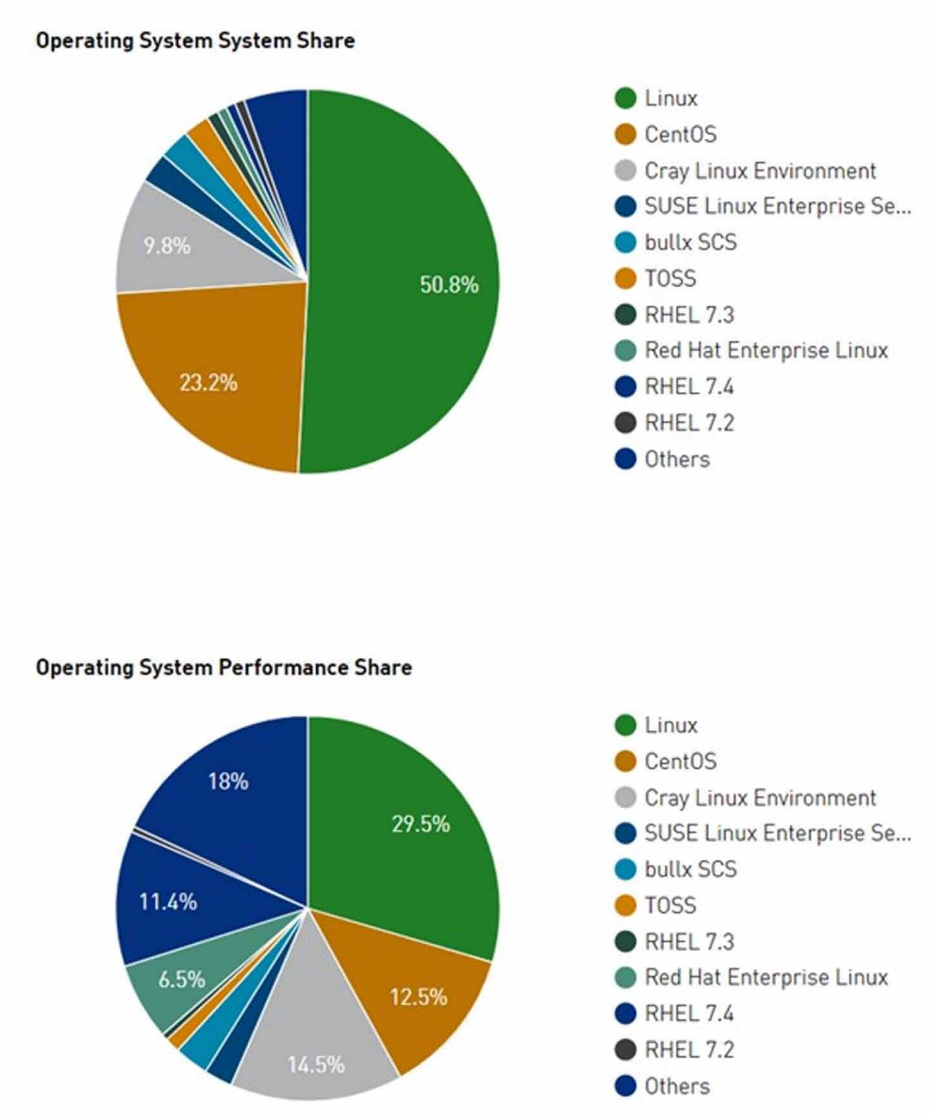
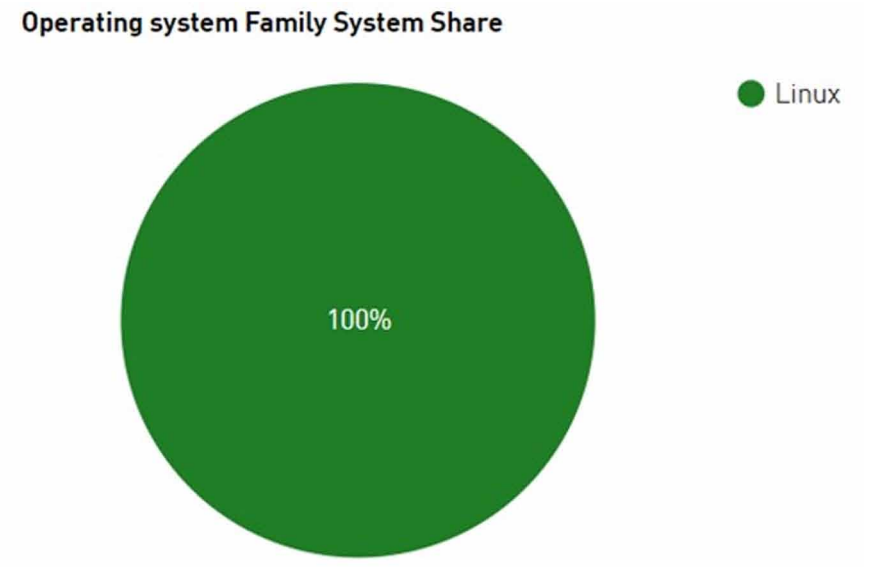


Table 12. Table of Top 500 Big Data storage servers list statistics by Operating System as of June 2018 (Top 500, 2018b). (Source: https://www.top500.org/statistics/list/)

Operating System	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
Linux	254	50.8	357,417,102	570,155,487	18,422,444
CentOS	116	23.2	151,128,500	279,748,037	7,971,198
Cray Linux Environment	49	9.8	175,908,272	274,713,912	6,482,376
SUSE Linux Enterprise Server 11	13	2.6	28,906,006	40,036,438	1,099,988
bullx SCS	13	2.6	35,160,329	56,897,152	1,332,400
TOSS	11	2.2	15,370,527	17,730,300	537,840
RHEL 7.3	5	1	6,488,070	11,031,808	136,656
Red Hat Enterprise Linux	4	0.8	78,835,380	129,943,142	1,696,896
RHEL 7.4	4	0.8	137,850,460	211,347,693	2,608,884
RHEL 7.2	4	0.8	5,736,500	6,562,202	174,208
Ubuntu Linux	4	0.8	6,981,000	14,878,722	177,704
SUSE Linux Enterprise Server 12	3	0.6	18,526,020	27,252,925	408,248
Bullx Linux	3	0.6	5,911,620	7,935,130	204,000
Kylin Linux	2	0.4	63,515,890	103,753,198	5,156,480
Redhat Enterprise Linux 6	2	0.4	2,433,470	3,032,783	295,656
Scientific Linux	2	0.4	2,889,600	4,626,616	67,704
Redhat Enterprise Linux 6.5	2	0.4	2,987,745	4,115,251	105,216
SLES12 SP2	1	0.2	3,328,590	4,777,574	78,336
Redhat Enterprise Linux 7	1	0.2	4,540,690	6,912,000	72,000
SUSE Linux	1	0.2	6,470,800	10,296,115	153,216
Sunway RaiseOS 2.0.5	1	0.2	93,014,594	125,435,904	10,649,600
RHEL 6.2	1	0.2	773,700	961,126	46,208
RHEL	1	0.2	1,022,000	1,484,000	17,640
Redhat Enterprise Linux 6.4	1	0.2	1,052,000	1,473,600	23,040
bullx SUperCOmputer Suite A.E.2.1	1	0.2	1,359,000	1,667,174	77,184
Ubuntu 14.04	1	0.2	3,307,000	4,896,512	60,512

Figure 13. Top 500 Big Data storage servers list statistics by Operating System Family as of June 2018 (Top 500, 2018b). (Source: https://www.top500.org/statistics/list/)



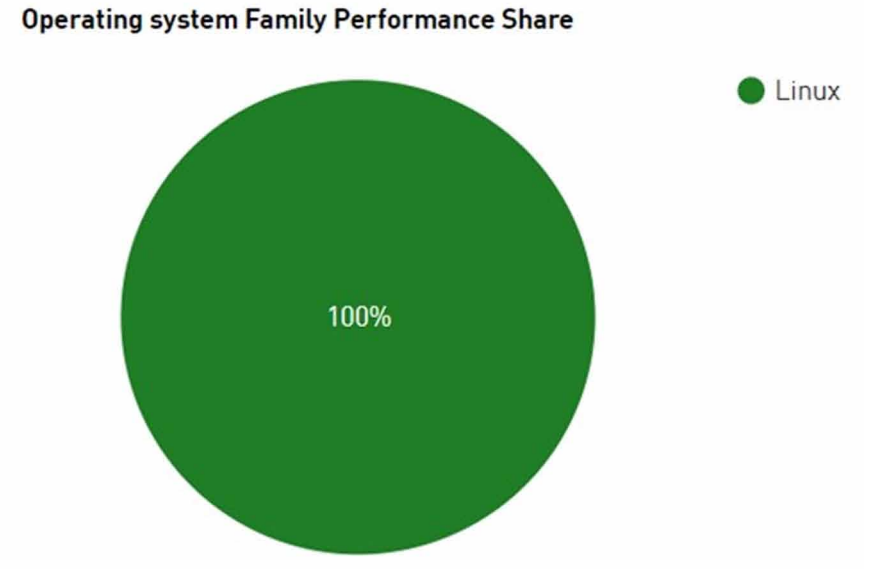


Table 13. Table of Top 500 Big Data storage servers list statistics by Operating System Family as of June 2018 (Top 500, 2018b). (Source: https://www.top500.org/statistics/list/)

Operating system Family	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
Linux	500	100	1,210,914,864	1,921,664,802	58,055,634

Figure 14. Top 500 Big Data storage servers list statistics by Accelerator/Co-Processor as of June 2018 (Top 500, 2018b). (Source: https://www.top500.org/statistics/list/)

Accelerator/Co-Processor	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
NVIDIA Tesla P100	54	10.8	92,474,566	157,608,577	2,141,088
NVIDIA Tesla V100	9	1.8	10,769,400	21,164,239	242,016
NVIDIA Tesla K80	7	1.4	10,834,746	17,986,269	341,390
NVIDIA Tesla K40	6	1.2	11,062,790	18,034,192	246,920
NVIDIA Tesla P100 NVLink	5	1	11,716,000	18,611,235	211,268
NVIDIA Tesla K20x	4	0.8	24,615,000	38,926,835	731,712
NVIDIA 2050	3	0.6	4,608,700	9,028,051	360,256
PEZY-SC2 500Mhz	3	0.6	2,480,317	3,337,933	2,351,424
Intel Xeon Phi 5110P	2	0.4	3,515,893	4,729,128	272,136
NVIDIA Volta GV100	2	0.4	193,910,000	306,852,868	3,855,024
NVIDIA Tesla P40	2	0.4	1,834,600	3,731,200	386,000
Intel Xeon Phi 5120D	2	0.4	2,571,200	4,099,204	122,512
NVIDIA Tesla K40m	1	0.2	2,478,000	4,946,790	64,384
NVIDIA Tesla K40/Intel Xeon Phi	1	0.2	3,126,240	5,610,481	152,692
NVIDIA Tesla V100 SXM2	1	0.2	19,880,000	32,576,635	391,680
Matrix-2000	1	0.2	61,444,500	100,678,664	4,981,760
Intel Xeon Phi 7110P	1	0.2	745,997	998,502	55,664
NVIDIA Tesla K20m	1	0.2	870,000	1,550,476	23,452
NVIDIA 2070	1	0.2	901,900	1,700,210	78,660
PEZY-SCnp	1	0.2	1,001,010	1,533,460	1,313,280
NVIDIA K20/K20x, Xeon Phi 5110P	1	0.2	1,018,000	1,502,236	222,072
Intel Xeon Phi 7120P	1	0.2	1,457,730	2,011,641	76,896
Intel Xeon Phi 31S1P	1	0.2	2,071,390	3,074,534	174,720

Table 14. Top 500 Big Data storage servers list statistics by Accelerator/Co-Processor (Top500, 2018b)

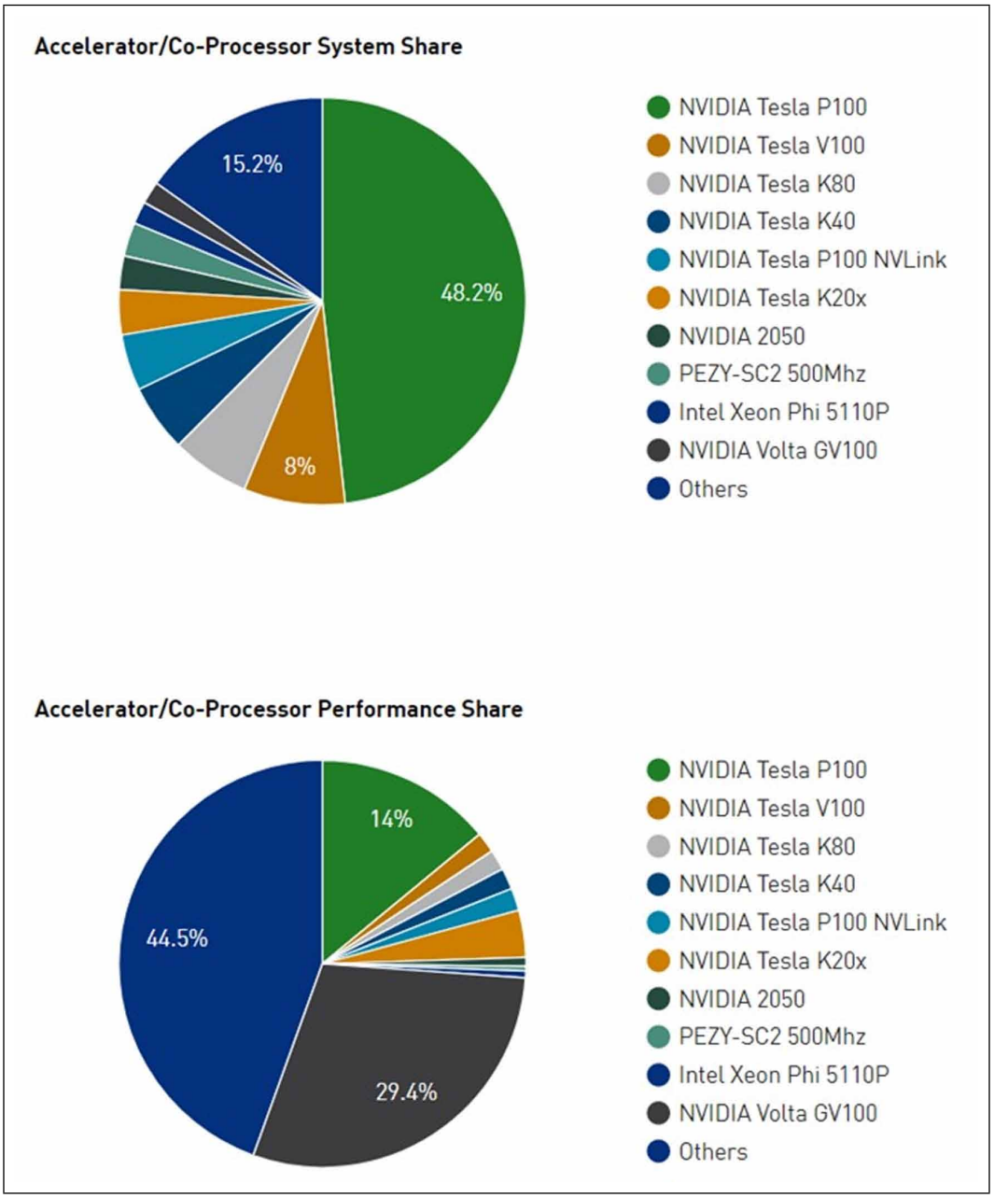
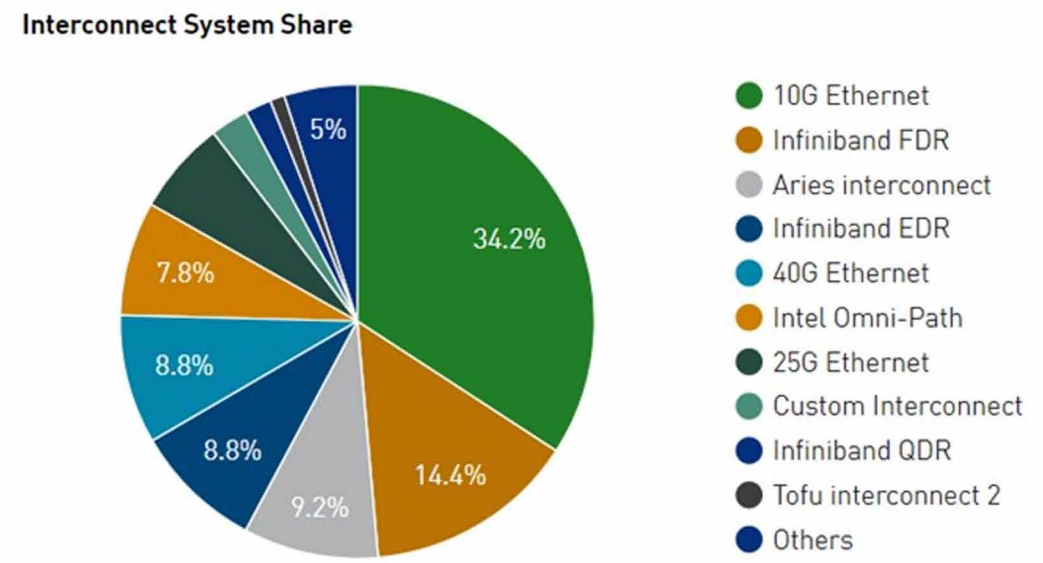


Figure 15. Top 500 Big Data storage servers list statistics by Interconnect as of June 2018 (Top 500, 2018b) (Source: https://www.top500.org/statistics/list/)



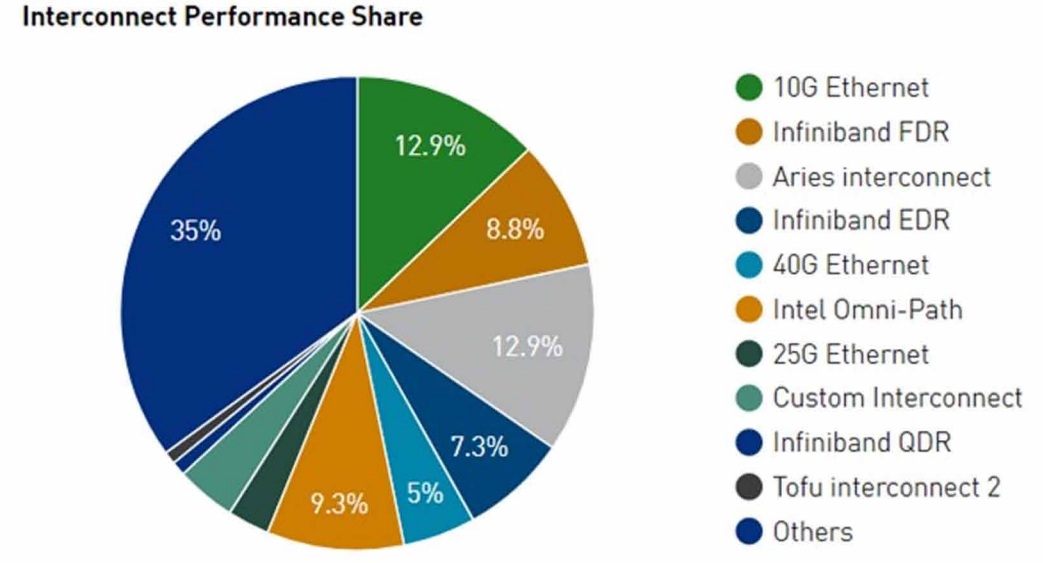


Table 15. Table of Top 500 Big Data storage servers list statistics by Interconnect as of June 2018 (Top 500, 2018b) (Source: https://www.top500.org/statistics/list/)

Interconnect	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
10G Ethernet	171	34.2	155,752,022	317,206,780	8,390,192
Infiniband FDR	72	14.4	107,003,766	153,310,488	5,343,984
Aries interconnect	46	9.2	156,425,782	244,823,186	5,844,416
Infiniband EDR	44	8.8	87,860,351	135,856,603	4,584,378
40G Ethernet	44	8.8	60,103,355	77,175,910	2,093,344
Intel Omni-Path	39	7.8	113,045,763	183,669,898	3,541,036
25G Ethernet	32	6.4	34,713,871	66,646,000	810,080
Custom Interconnect	13	2.6	49,484,541	57,042,487	4,192,224
Infiniband QDR	9	1.8	11,834,759	18,615,972	740,584
Tofu interconnect 2	5	1	10,422,200	11,530,310	354,384
Bull BXI 1.2	4	0.8	15,246,719	28,804,115	670,712
Mellanox InfiniBand EDR	3	0.6	20,816,240	32,168,658	405,088
Dual-rail Mellanox EDR Infiniband	3	0.6	194,928,000	308,338,395	3,874,464
Infiniband	3	0.6	2,650,905	4,187,301	146,244
Cray Gemini interconnect	2	0.4	18,757,000	28,617,830	711,168
TH Express-2	2	0.4	63,515,890	103,753,198	5,156,480
Infiniband EDR/FDR	2	0.4	1,996,336	3,596,544	74,160
Proprietary	2	0.4	3,337,700	6,043,751	239,616
Mellanox EDR InfiniBand/ParTec ParaStation	1	0.2	6,177,730	9,891,072	114,480
Sunway	1	0.2	93,014,594	125,435,904	10,649,600
56G Infiniband FDR	1	0.2	1,013,721	1,372,134	32,984
Infiniband FDR14	1	0.2	2,813,620	3,578,266	86,016

Figure 16. Real-time Big Data Processing with I/O Centric Storage (Floyer, 2012)

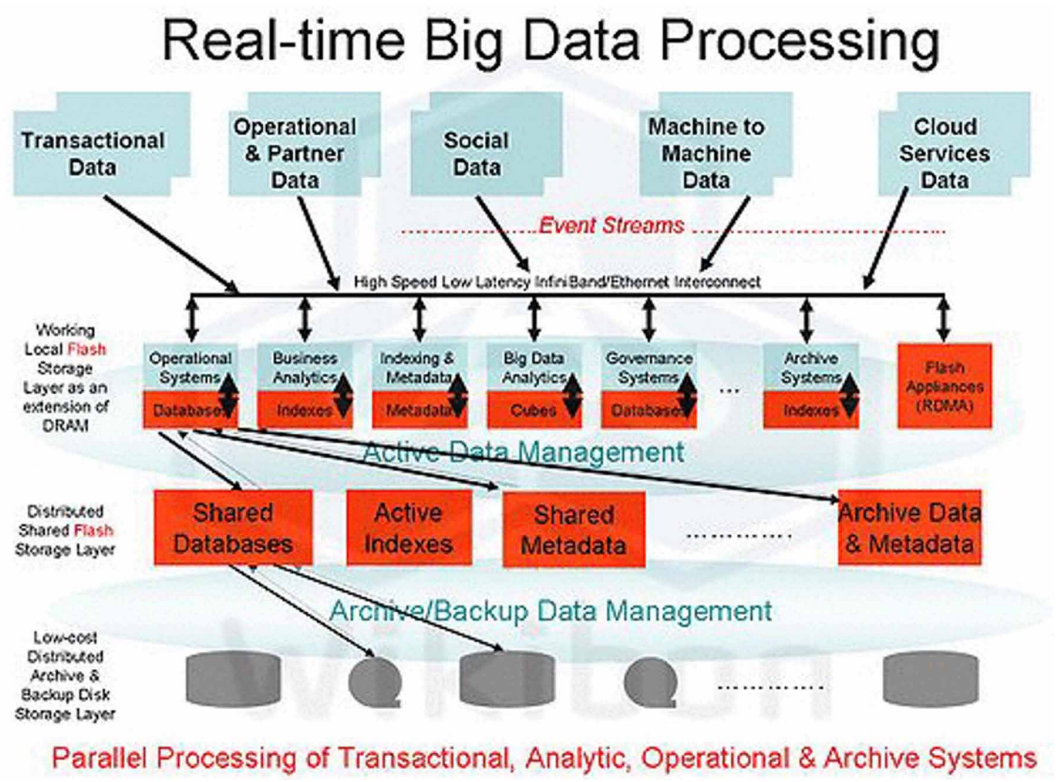
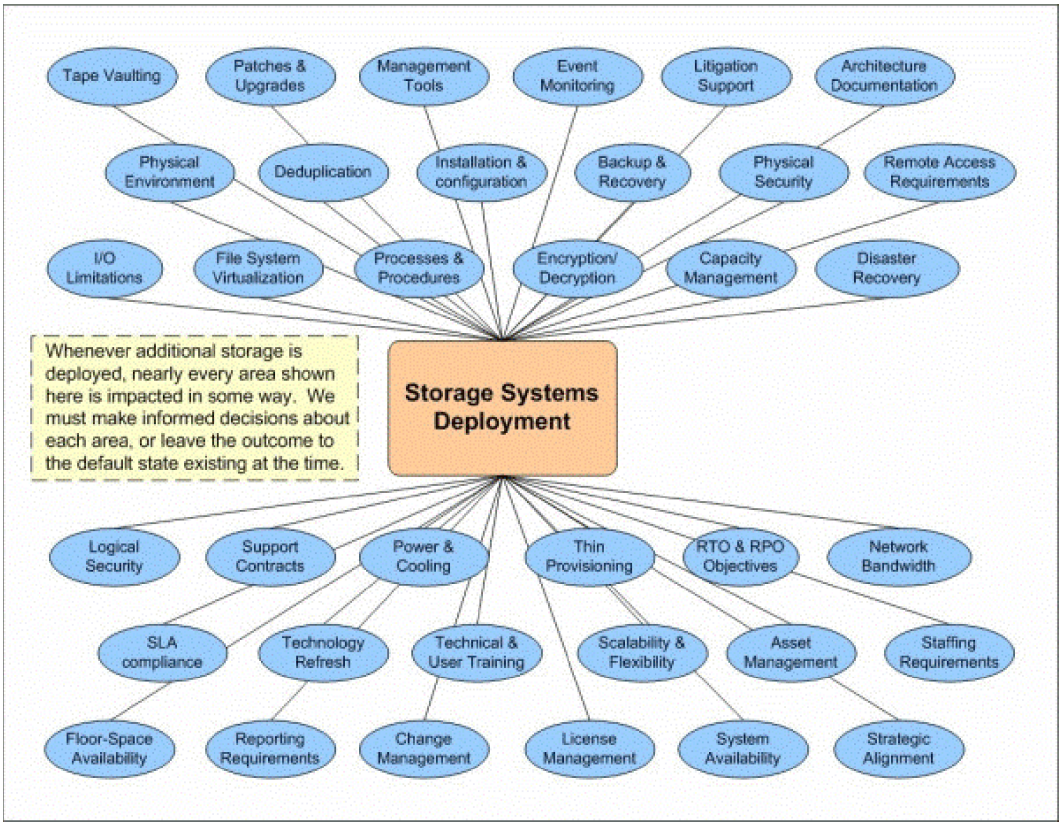


Figure 17. Areas that are directly or indirectly affected by storage growth ("Big Data" Getting Bigger? Beware of the Ripple Effect, (Cochran, 2012)).



REFERENCES

Achahbar, O., & Abid, M. R. (2015). The impact of virtualization on high performance computing clustering in the cloud. *International Journal of Distributed Systems and Technologies, 6*(4), 65-81. Retrieved from https://www.researchgate.net/publication/282531800_The_Impact_of_Virtualization_on_High_Performance_Computing_Clustering_in_the_Cloud

Ahuja, S. P., & Deval, N. (2018). From Cloud Computing to Fog Computing: Platforms for the Internet of Things (IoT). *International Journal of Fog Computing*, *1*(1), 1–14. doi:10.4018/IJFC.2018010101

Alageswaran, R., & Amili, A. M. J. (2018). Evolution of Fog Computing and Its Role in IoT Applications. In P. Raj & A. Raman (Eds.), Handbook of Research on Cloud and Fog Computing Infrastructures for Data Science (Ch. 2, pp. 33-52). Hershey, PA: IGI Global.

Alonso-Monsalve, S., Garcia-Carballeria, F.-G., & Calderon, A. (2017). Fog computing through public-resource computing and storage. In *Proceedings of 2017 Second International Conference on Fog and Mobile Edge Computing (FMEC)*, May 8-11. doi:10.1109/FMEC.2017.7946412

Azeem, S. A., & Sharma, S. K. (2016). Study of converged infrastructure & hyper converge infrastructre as future of data centre. *International Journal of Advanced Research in Computer Science*. Retrieved from http://www.ijarcs.info/index.php/Ijarcs/article/view/3476

Balakrishnan, P., Venkatesh, V., & Raj, P. (2018). Fog Computing: Introduction, Architecture, Analytics, and Platforms. In P. Raj & A. Raman (Eds.), Handbook of Research on Cloud and Fog Computing Infrastructures for Data Science (Ch. 2, pp. 68-84). Hershey, PA: IGI Global.

Barney, B. (2017). *Message Passing Interface (MPI)*. U.S. Department of Energy (DOE), Lawrence Livermore National Laboratory (LLNL). Retrieved from https://computing.llnl.gov/tutorials/mpi/

Beaver, D., Kumar, S., Li, H. C., Sobel, J., & Vajget, P. (2010). Finding a needle in a haystack: Facebook's photo storage. In *Proceedings of the Ninth USENIX Conference on Operating Systems Design and Implementation* (pp. 1-8). Berkeley: CA, USENIX Association. Retrieved from https://www.usenix.org/legacy/event/osdi10/tech/full_papers/Beaver.pdf

Belli, L., Cirani, S., Davoli, L., Ferrari, G., Melegari, L., Montón, M., & Marco Picone, M. (2018). A Scalable Big Stream Cloud Architecture for the Internet of Things. In Fog Computing: Breakthroughs in Research and Practice (Ch. 2, pp. 25-53). Hershey, PA: IGI Global.

Bhatt, C., & Bhensdadia, C. K. (2018). Fog Computing: Applications, Concepts, and Issues. In Fog Computing: Breakthroughs in Research and Practice (Ch. 9, pp. 198-207). Hershey, PA: IGI Global.

Butler, B. (2013, January 4). Top 10 cloud storage providers according to Gartner, *ComputerWorldUK*. Retrieved from http://www.computerworlduk.com/it-vendors/top-10-cloud-storage-providers-according-gartner-3418594/

Carpenter, J., & Hewitt, E. (2016). *Cassandra the definite guide: Distributed data at web scale* (2nd ed.). Sebastopol, CA: O'Reilly Media, Inc. Retrieved from http://shop.oreilly.com/product/0636920043041.do

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., & Burrows, M. … Gruber, R.E. (2008). BigTable: A distributed storage system for unstructured data. *ACM Transactions on Computer Systems*, 26(2), 4. Retrieved from https://static.googleusercontent.com/media/research.google.com/en//archive/bigtable-osdi06.pdf

Chen, J., Choudhary, A., Feldman, S., Hendrickson, B., Johnson, C., & Mount, R. … Williams. D. (2013). Synergistic challenges in data-intensive science and exascale computing. *US Department of Energy (DOE) Advanced Scientific Computing Advisory Committee (ASCAC)*. Retrieved from http://science.energy.gov/~/media/40749FD92B58438594256267425C4AD1.ashx

Chen, M., Mao, S., Zhang, Y., & Leung, V. C. M. (2014). *Big Data: Related technologies, challenges and future prospects*. Springer. Retrieved from http://www.springer.com/us/book/9783319062440

Chiang, M., & Zhang, C. (2016). Fog and IoT: An Overview of Research Opportunities. *IEEE Internet of Things Journal*, *3*(6), 854-864. doi: Retrieved from http://www.download-paper.com/wp-content/uploads/2017/01/2016-ieee-Fog-and-IoT-An-Overview-of-Research-Opportunities.pdf10.1109/EuCNC.2017.7980667

Cisco. (2015). White Paper: Fog computing and the Internet of Things: Extend the cloud to where the things are. Retrieved from https://www.cisco.com/c/dam/en_us/solutions/trends/iot/docs/computing-overview.pdf

Cochran, R. (2012). *Big data getting bigger? Beware of the ripple effect.* Big Data Challenges. Data Center Enhancements Inc., Retrieved from http://bigdatachallenges.com/2012/03/02/big-data-getting-bigger-beware-of-the-ripple-effect/

COMSOL, Inc. (2017). High Performance Computing (HPC). *Multiphysics CLOPEDIA*. Retrieved from https://www.comsol.com/multiphysics/high-performance-computing

Data-Intensive Scalable Computing Laboratory (DISCL). (2015). Retrieved from http://discl.cs.ttu.edu/doku.php?id=projects

DataDirect Networks. (2011). DataDirect Networks' (DDN) big data storage technology powers more than 60 percent of the world's 100 fastest computers Retrieved from http://www.ddn.com/press-releases/datadirect-networks-ddn-big-data-storage-technology-powers-60-percent-worlds-100-fastest-computers/

Deka, G. C. (2017). *NoSQL: Database for storage and retrieval of data in cloud* (1st ed.). Boca Raton, FL: Chapman and Hall/CRC. Retrieved from https://www.amazon.com/NoSQL-Database-Storage-Retrieval-Cloud-ebook/dp/B072BZ5D4T/ref=sr_1_1?s=books&ie=UTF8&qid=1501446340&sr=1-1&keywords=deka+No+SQL+Database+for+Storage

Estrada, R., & Ruiz, I. (2016). *Big data SMACK: A guide to Appache Spark, Mesos, Akka, Cassandra, and Kafka*. New York, NY: Springer. Retrieved from http://www.apress.com/us/book/9781484221747

Floyer, D. (2012). *Assessment of EMC Project Thunder, Server Area Networks*. Retrieved from http://wikibon.org/wiki/v/Assessment_of_EMC_Project_Thunder,_Server_Area_Networks

Gadepally, V., Kepner, J., & Reuther, A. (2016). *Storage and database management for big data*, Chapter 2 of Big Data: Storage, Sharing and Security, Edited by Hu, F. (2016), CRC Press, Boca Raton: FL, pp. 15-42. Retrieved from https://www.crcpress.com/Big-Data-Storage-Sharing-and-Security/Hu/p/book/9781498734868

Gao, X., Roth, E., McKelvey, K., Davis, C., Younge, A., Ferrara, E., . . . Qiu, J. (2014). Supporting a social media observatory with customizable index structure: architecture and performance, In *Cloud Computing for Data-Intensive Applications* (pp. 401-427). New York, NY: Springer Science+Business Media. Retrieved from http://www.springer.com/us/book/9781493919048

Gartner. (2012). *Gartner says public cloud services are simultaneously cannibalizing and stimulating demand for external IT services spending*. Retrieved from http://www.gartner.com/newsroom/id/2220715

Grieco, C. (2017). *Spark™ big data cluster computing in production*. CreateSpace Independent Publishing Platform. Retrieved from http://www.amazon.in/Spark-Data-Cluster-Computing-Production/dp/1119254019

Gupta, N. (2015). Top500 supercomputers in the world. In R.S. Segall, J.S. Cook, & Q. Zhang (Eds.), *Research and Applications in Global Supercomputing* (pp. 445-588). Hershey, PA: IGI Global. Retrieved from https://www.igi-global.com/book/research-applications-global-supercomputing/118093

Hosken, M. (2016). *VMware software-defined storage: A design guide to the policy-driven, software-defined storage era* (1st ed.). Sybex Publishing. Retrieved from http://www.wiley.com/WileyCDA/WileyTitle/productCd-1119292778,miniSiteCd-SYBEX.html

Hu, F. (Ed.). (2016). *Big data: Storage, sharing and security*. Boca Raton, FL: CRC Press. Retrieved from https://www.crcpress.com/Big-Data-Storage-Sharing-and-Security/Hu/p/book/9781498734868

IBM. (2017). *HPSS: High Performance Storage System*. Retrieved from http://www.hpss-collaboration.org/

Icon Group International. (2018). *The 2018-2023 world outlook for big data storage*. San Diego, CA: ICON Group International, Inc. Retrieved from https://www.amazon.com/s/ref=nb_sb_noss?url=search-alias%3Dstripbooks&field-keywords=the+2018-2023+world+outlook+for+big+data+storage

Intel White Paper. (2014). *Big data meets high performance computing*. Retrieved from http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/big-data-meets-high-performance-computing-white-paper.pdf

Jackson, J. C., Vijayakumar, V., Quadir, M. A., & Bharathi, C. (2015). Survey on programming models and environments for cluster, cloud, and grid computing that defends big data. *Procedia Computer Science*, *50*, 517–523. doi:10.1016/j.procs.2015.04.025

Kleppman, M. (2017). *Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems.* Sebastopol, CA: O'Reilly Media, Inc. Retrieved from http://shop.oreilly.com/product/0636920032175.do

Li, X., & Qiu, J. (Eds.). (2014). *Cloud computing for data-intensive applications.* Springer Science+ Business Media. Retrieved from http://www.springer.com/us/book/9781493919048

Ostberg, P.-O. (2017). Reliable Capacity Provisioning for Distributed Cloud/Edge/Fog Computing Applications. In *Proceedings of 2017 European Conference on Networks and Communications (EuCNC)*. doi: Retrieved from http://eprints.networks.imdea.org/1625/1/1570343878.pdf10.1109/JIOT.2016.2584538

Perera, C., Qin, Y., Estrella, J. C., Reiff-Marganiec, S., & Vasilakos, A. V. (2017). Fog Computing for Sustainable Smart Cities: A Survey. [CSUR]. *ACM Computing Surveys*, *50*(3), 32. doi:10.1145/3057266

Pierson, F. (2017, July 13). Breaking down the 4 of the best big data filesystems. *Big Data Zone*. Retrieved 2017 from https://dzone.com/articles/breaking-down-the-4-of-the-best-big-data-filesyste

Qiang, W., Zheng, X., & Hsu, C.-H. (2016). Cloud computing and big data. In *Second international conference, CloudCom-Asia 2015*, Huangshan, China, June 17-19. Springer International Publishing, Switzerland. ISBN 978-3-319-28429-3. Retrieved on August 3, 2017 from https://www.amazon.com/gp/product/3319284290/ref=oh_aui_detailpage_o00_s00?ie=UTF8&psc=1

Radadiya, M., & Rohokale, V. (2016). Implementation of costing model for high performance computing as a services on the cloud environment. In *AICTC '16 Proceedings of the International Conference on Advances in Information Communication Technology & Computing*. Bikaner, India, August 12-13. ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2979841

Raj, P., & Raman, A. (2018). *Handbook of Research on Cloud and Fog Computing Infrastructures for Data Science*. Hershey, PA: IGI Global. doi:10.4018/978-1-5225-5972-6

Ramakrishnan, L., Ghoshal, D., Hendrix, V., Feller, E., Mantha, P., & Morin, C. (2017). Storage and Data Life Cycle Management in Cloud Experiments with FRIEDA. In *Cloud Computing for Data-Intensive Applications* (pp. 357-378). Retrieved from https://link.springer.com/chapter/10.1007/978-1-4939-1905-5_15

Reddy, S., & Raz, J. (2017). Hosting and delivering Casandra NoSQL database via cloud environments. In Deka, G.C. (2017). *NoSQL: Database for Storage and Retrieval of Data in Cloud* (1st ed.). Boca Raton, FL: Chapman and Hall/CRC. Retrieved from https://www.crcpress.com/NoSQL-Database-for-Storage-and-Retrieval-of-Data-in-Cloud/Deka/p/book/9781498784368

Robb, D. (2016). Top ten big data storage tools. *Infostor*. Retrieved from http://www.infostor.com/backup-and_recovery/top-ten-big-data-storage-tools.html

Ross, B., Arslan, E., Zhang, B., & Kosar, T. (2017). Managed file transfer as a cloud service. In *Cloud computing for data-intensive applications* (pp. 379-400). Retrieved from https://link.springer.com/chapter/10.1007%2F978-1-4939-1905-5_16

Rouse, M. (2017a). Apache Hadoop YARN (Yet Another Resource Negotiator). *Search Data Management*. Retrieved from http://searchdatamanagement.techtarget.com/definition/Apache-Hadoop-YARN-Yet-Another-Resource-Negotiator

Rouse, M. (2017b). High Performance Computing (HPC). *TechTarget.* Retrieved from http://searchdatacenter.techtarget.com/definition/high-performance-computing-HPC

Rouse, M. (2017c). Message passing interface (MPI). *TechTarget.* Retrieved from http://searchenterprisedesktop.techtarget.com/definition/message-passing-interface-MPI

Sakr, S. (2016). Big data 2.0 processing systems: A survey. Springer. ISBN 978-3-319-38775-8 (Print) and ISBN 978-3-319-38776-5 (eBook) SpringerBriefs in Computer Scence, Springer, Switzerland. https://www.springer.com/us/book/9783319387758

Sandisk White Paper. (2015). *Storage in the era of cloud and big data: the advantages of SSDs over HDDs*. Retrieved from http://www.sandisk.com/assets/docs/WP009_White%20Paper%20-%20Storage%20in%20the%20Era%20of%20Cloud%20and%20Big%20Data%20-%20the%20Advantages%20of%20SSDs%20over%20HDDs%20-%2012202013%20FINAL.pdf

Sawant, N., & Shah, H. (2013). *Big data application architecture Q&A: A problem-solution approach*. Apress Media, LLC. Retrieved from http://www.apress.com/us/book/9781430262923

Seagate. (2015). *The art of high performance scale-out storage.* Retrieved from http://www.seagate.com/products/enterprise-servers-storage/enterprise-storage-systems/clustered-file-systems/

Segall, R. S. (2013). Computational Dimensionalities of Global Supercomputing. *Journal of Systemics, Cybernetics and Informatics*, *11*(9), 75–86. Retrieved from http://www.iiisci.org/journal/sci/FullText.asp?var=&id=iSA625MW

Segall, R. S. (2015). Invited Plenary Address at International Institute of Informatics and Systemics (IIIS) Conference titled: "Research and Applications in Global Supercomputing: An Interdisciplinary Science". In 18th Multi-conference on Systemics, Cybernetics, and Informatics (WMSCI 2014), Orlando, FL, July 15-18. Retrieved from http://www.iiis.org/ViewVideo2014.asp?id=10

Segall, R.S., Cook. J.S., & Zhang, Q. (Eds.) (2015). *Research and applications in global supercomputing*. Hershey, PA: IGI Global. ISBN 13: 978-1-46-667461-5. Retrieved from https://www.igi-global.com/book/research-applications-global-supercomputing/118093

Segall, R. S. (2016a). Invited Plenary Address at International Institute of Informatics and Systemics (IIIS) Conference titled: "Big Data: A Treasure Chest for Interdisciplinary Research".. In *20th Multi-conference on Systemics, Cybernetics, and Informatics (WMSCI 2016)*, Orlando, FL, July 5-8. Retrieved from http://www.iiis.org/ViewVideo2016.asp?id=14

Segall, R. S. (2016b). High performance computing and data mining in bioinformatics. In *13th Annual Meeting of MidSouth Computational Biology and Bioinformatics Society (MCBIOS)*, Memphis, TN, March 3-5.

Segall, R. S. (2017a). Using Tablets and Mobile Devices for Visual Analytics of Big Data in Bioinformatics. *Presentation at 14th Annual Meeting of MidSouth Computational Biology and Bioinformatics Society (MCBIOS)*, Little Rock, AR, March 23-25. Retrieved from https://mcbios.org/sites/mcbios.org/files/MCBIOS2017_Program_Book_Final_1_0.pdf

Segall, R. S. (2017b). Technologies for Teaching Big Data Analytics. In *Proceedings of 48th Meeting of Southwest Decision Sciences (SWDSI)*, Little Rock, AR, March 8-11. Retrieved from http://www.swdsi.org/swdsi2017/SWDSI_2017_CONFERENCE_PROGRAM4.pdf

Segall, R. S., & Gupta, N. (2015). Overview of global supercomputing. Chapter 1 in *Research and Applications in Global Supercomputing,* pp. 1-32. Hershey, PA: IGI Global. Retrieved from https://www.igi-global.com/chapter/overview-of-global-supercomputing/124335

Segall, R. S., & Niu, G. (2018). Overview of Big Data and Its Visualization. Chapter 1 in Handbook of Big Data Storage and Visualization Techniques. Hershey, PA: IGI Global.

Segall, R.S. & Niu, G. (2018). Overview of Big Data and its Visualizations with Fog Computing. *International Journal of Fog Computing*. ,Vol. 1, No.2, pp. 51-82. Available at https://www.igi-global.com/article/big-data-and-its-visualization-with-fog-computing/210566

Swami, D., Sahoo, S., & Sahoo, B. (2018). Storing and analyzing streaming data: A big data challenge. In Big Data Analytics: Tools and Technology for Effective Planning (pp. 229-246). Boca Raton, FL: CRC Press. Available at https://www.crcpress.com/Big-Data-Analytics-Tools-and-Technology-for-Effective-Planning/Somani-Deka/p/book/9781138032392. (Edited by A.K. Somani and G.C. Deka),

Techopedia, Inc. (2017). Amdahl's Law. Retrieved from https://www.techopedia.com/definition/17035/amdahls-law

Top 500. (2018a). Top 10 Sites for June 2018. Retrieved from https://www.top500.org/lists/2018/06/

Top 500. (2018b). *List Statistics.* Retrieved from https://www.top500.org/statistics/list/

Top 500. (2018c). Treemaps. Retrieved from https://www.top500.org/statistics/treemaps/

Top 500. (2018d). Efficiency Power Cores. Retrieved from https://www.top500.org/statistics/efficiency-power-cores/

Top 500. (2018e). Development Over Time. Retrieved from https://www.top500.org/statistics/overtime/

Tudoran, R., Costan, A., Antoniu, G., & Goetz, B. (2014). Big data storage and processing on Azure clouds: Experiments on scale and lessons learned. In *Cloud Computing for Data-Intensive Applications* (pp. 331-356). New York, NY: Springer Science+Business Media. Retrieved from http://www.springer.com/us/book/9781493919048

WhoIsHostingThis.com. (2017). *MPI - Introduction, history and resources.* Retrieved from http://www.whoishostingthis.com/resources/mpi/#reference

Wikipedia. (n.d.). Data-intensive computing. Retrieved July 28, 2017 from http://en.wikipedia.org/wiki/Data-intensive_computing

Wikipedia. (n.d.). InfiniBand (IB). Retrieved August 3, 2017 from https://en.wikipedia.org/wiki/InfiniBand

Wikipedia. (n.d.). Network File System. Retrieved August 2, 2017 from https://en.wikipedia.org/wiki/Network_File_System

Wikipedia. (n.d.). Parallel Virtual File System (PVFS). Retrieved August 3, 2017 from https://en.wikipedia.org/wiki/Parallel_Virtual_File_System

Wikipedia. (n.d.). *Platform-as-a-Service (PaaS).* Retrieved August 3, 2017 from https://en.wikipedia.org/wiki/Platform_as_a_service

Winn, M., Follows, J., Rawlings, C., Caccamo, M., & Flicek, P. (2012). Data -intensive computing in biology. *CECAM.org*. Retrieved from http://www.cecam.org/workshop-726.html

Wu, D., Sakr, S., & Zhu, L. (2017a). Big data storage and data models. In Handbook of Big Data Technologies (pp. 3-29). Springer International Publishing. doi:10.1007/978-3-319-49340-4_1

Wu, D., Sakr, S., & Zhu, L. (2017b) Big data programming models, In Handbook of Big Data Technologies (pp. 31-63). Springer International Publishing. doi:10.1007/978-3-319-49340-4_2

Xu, Y. (2016). *Storage management of data-intensive computing systems*. Florida International University. Retrieved from http://digitalcommons.fiu.edu/etd/2474

Zverina, J. (2012). SDSC supercharges its 'data oasis' storage system. *San Diego Supercomputing Center (SDSC)*. Retrieved from http://ucsdnews.ucsd.edu/pressrelease/sdsc_supercharges_its_data_oasis_storage_system

Richard S. Segall is Professor of Computer & Information Technology in Neil Griffin College of Business at Arkansas State University. He holds BS/MS in mathematics, MS in operations research and statistics from Rensselaer Polytechnic Institute in Troy, New York, and PhD in operations research form University of Massachusetts at Amherst. He has served on faculty of Texas Tech University, University of Louisville, University of New Hampshire, University of Massachusetts-Lowell, and West Virginia University. His research interests include data mining, Big Data, text mining, web mining, database management, and mathematical modeling. His funded research includes that by U.S. Air Force, NASA, Arkansas Biosciences Institute (ABI), and Arkansas Science & Technology Authority (ASTA), and is member of the Arkansas Center for Plant-Powered-Production (P3), Editorial Board of International Journal of Data Mining, Modelling and Management (IJDMMM), International Journal of Data Science (IJDS), and co-editor of Handbook of Big Data Storage and Visualization Techniques; Research and Applications in Global Supercomputing; Visual Analytics of Interactive Technologies: Applications to Data, Text & Web Mining, all published by IGI Global.

Jeffrey S. Cook lives in Paragould, Arkansas, where he is an entrepreneur in computer technology and artificial technology. He was co-editor of book Research and Applications in Global Supercomputing edited by Richard S. Segall, Jeffrey S. Cook and Qingyu Zhang that was published by IGI Global in 2015. He is co-author of articles: "Data Visualization and Information Quality by Supercomputing" by Richard S. Segall and Jeffrey Cook that was published in the Proceedings of the Forty-Fifth Meeting of Southwest Decision Sciences Institute (SWDSI) that was held in Dallas, TX in March 12-15, 2014; and "Overview of Current Research in Global Supercomputing" authored by Richard S. Segall, Qingyu Zhang, and Jeffrey S. Cook that was published in Proceedings of Forty-Fourth Meeting of Southwest Decision Sciences Institute (SWDSI), that was held in Albuquerque, NM, March 12-16, 2013. He has also contributed a chapter entitled "Supercomputers and Supercomputing" in Visual Analytics and Interactive Technologies: Data, Text and Web Mining Applications edited by Qingyu Zhang, Richard Segall, and Mei Cao that was published by IGI Global in 2011. He has also presented a workshop on Bioinformatics Tools at the 2010 MidSouth Computational Biology and Bioinformatics Society (MCBIOS) Annual meeting that was held at Arkansas State University. He currently holds a patent in Aerospace-design that was adopted by Boeing Aerospace Corporation along with several other patents and sub-patents pertaining to software development. He is also a research writer who has published several books through McGraw-Hill College Division and Vantage Press. Jeffrey S. Cook is listed in Cambridge's book of Who's Who for most influential people in the world.

Gao Niu is an Assistant Professor in Actuarial Science at Bryant University. He also serves as the Assistant Director of the Janet & Mark L Goldenson Center for Actuarial Research at the University of Connecticut. He has a doctorate in actuarial science from the University of Connecticut, is an Associate of the Casualty Actuarial Society and a Member of the American Academy of Actuaries. Dr. Niu has years of experience in academic actuarial research and consulting practice. His research area includes but not limited to the following: big data analytics application in insurance industry, property and casualty insurance practice, predictive modeling, agent based modeling, financial planning, life insurance and health insurance pricing, reserving and data mining.