WHITE PAPER

Big Data and the Network

Sponsored by: Brocade/EMC Greenplum

Richard L. Villars Lucinda Borovick

November 2011

IDC OPINION

In the past, the main data challenge for most organizations was enabling/recording more transactions more quickly. Today, much of the focus is on faster delivery of more information from scale-out cloud computing clusters (e.g., documents, medical images, movies, gene sequences, data streams, tweets) to systems, PCs, mobile devices, and living rooms. The challenge going forward will be finding ways to better analyze, monetize, and create new value from all this information. Welcome to the era of Big Data.

Big Data describes a new generation of technologies and architectures designed so that organizations across many different industries and sectors can economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and analysis. Big Data requires a shift in computing, storage, and networking architecture so that organizations can handle the ingest, management, and distributed processing workloads required to analyze large volumes of data economically.

IDC believes the organizations that are best able to make real-time business decisions using Big Data streams will thrive, while those that are unable to embrace and make use of this shift will increasingly find themselves at a competitive disadvantage in the market and face potential failure. This will be particularly true in industries experiencing high rates of business change and aggressive consolidation.

The network is the supporting foundation necessary for successful Big Data deployments and an organization's ability to leverage massive amounts of diverse data sources. IDC looks for network attributes that incorporate automation, scalability, and high performance onto an agile datacenter fabric to meet the requirements for emerging Big Data deployments.

INFORMATION EVERYWHERE, BUT WHERE'S THE KNOWLEDGE?

Explosive data growth is a reality. The rapid acceleration of data volume growth will persist in the foreseeable future, thanks to the proliferation of new computing and communication devices such as smartphones, sensors, and RFIDs and the continuously expanding need for business efficiency, innovation, and competitive advantage through better information management and advanced analytics. These applications allow an organization to interact and communicate with end users in both consumer and business environments. Enterprises and other organizations are:

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

- Collecting, storing, and analyzing more granular information about more products, people, and transactions than ever before.

- Relying on email, collaboration tools, and mobile devices to communicate and conduct business with customers and business partners.

- Creating, receiving, and collecting machine- and sensor-generated messages, sometimes at very high volumes, and are driving operations and business processes from that message data.

IDC's research puts the pace and newness of this data explosion into context. In 2007, organizations around the world installed 6.1 exabytes (EB) of disk storage capacity, mainly to store transaction data. By 2010, annual new installations were 16.4EB, with storage of rich content driving much of the demand. By 2014, new capacity installed will reach 79.8EB. Active and ongoing analysis of new data sources will be the most important force driving the next wave of datacenter expansion.

This massive data pool creates new data ingest, aggregation, and analytic challenges that require expanded use of multicore server architectures, active archival storage systems, and very high speed/high capacity networks — all deployed at sometimes hard-to-imagine scale The era of Big Data imposes significant new demands on the network, both in the datacenter and across the wide area network.

WHAT IS BIG DATA?

Big Data is about the volume, variety, and velocity of information being generated today and the opportunity that results from effectively leveraging data for insight and competitive advantage. Big Data describes a new generation of technologies and architectures designed to economically extract value from these very large and diverse volumes of data by enabling high-velocity capture, discovery, and/or analysis.

The Big Data trend is relevant to organizations across industry sectors worldwide. Industries that recently began to digitize their business are at the forefront of this development. In virtually all of these cases, data growth rates in the past five years have been near infinite because, in most cases, they started from zero. Industries tend to cluster into two groups:

- Large content creators/distributors

  - **Media/Web:** Collecting large amounts of rich content and user viewing behaviors

  - **Healthcare:** Storing electronic medical records, test results, and images for short-term public health monitoring and long-term epidemiological research

  - **Life sciences:** Looking for genetic variations and potential treatment effectiveness through low-cost gene sequencing (<$1,000)

  - **Video surveillance:** Analyzing behavioral patterns (security and service enhancement) based on high-definition IPTV output

- Machine-generated data creators/users

  - **Transportation/logistics:** Collecting/tracking RFID data to better manage inventories and detect smuggling

  - **Retail**: Tracking in-store, Web site, and social media transactions/comments to identify new opportunities and manage brand image

  - **Utilities:** Collecting/analyzing sensor data in real time to enable smart grid energy management systems

  - **Telecommunications:** Leveraging call data records (CDRs) and geographic location data to detect fraud and reduce churn

All industries and sectors depend on rapid and reliable data between the data sources (e.g., sensors, OLTP systems, cameras) and the analytic grid as well as highly scalable interconnect between the nodes of the grid. The ultimate value of a Big Data implementation will be judged based on one or more of three criteria:

- **Does it provide more useful information?** For example, a major retailer might link a digital video system in its stores to a social media analytic data source to analyze the flow of shoppers — including demographic information such as gender and age — through the store at different times of the day, week, and year. It could also compare flows in different regions with core customer demographics. This move makes it easier for the retailer to tune layouts and promotion spaces on a store-by-store basis.

- **Does it improve the timeliness of the response?** For example, several private and government healthcare agencies around the world are deploying Big Data systems to reduce the time to detect insurance fraud from months (after checks have been mailed and cashed) to days (eliminating the legal and financial costs associated with fund recovery).

- **Does it improve the fidelity of the information?** For example, energy companies and earth/life science research teams want to use Big Data systems to monitor and assess the quality of data being collected from remote sensor systems; they are using Big Data not just to look for patterns but also to identify and eliminate false data caused by malfunctions, user error, or temporary environmental anomalies.

BIG DATA DEPLOYMENT: A PRACTICAL GUIDE

Today, Big Data initiatives in many organizations are best described as "junior science projects" with a small core of servers and storage assets. They aren't the next iteration of a Google-like compute grid, at least not yet. From a business and an IT governance standpoint, however, these kinds of "junior science projects" can quickly turn into the next "Manhattan project" with companywide and industrywide business, organizational, and legal consequences.

As an organization makes the transition from Big Data as a "junior science project" to Big Data as a core business resource, concerns about the impact on current and future datacenters will increase. Today, and for quite some time, the IT architectural approach used in clustered environments such as a large Hadoop grid is radically different from the converged and virtualized IT environments driving most organizations' datacenter transformation strategies. They have different server/storage/network configurations and different environmental (power and HVAC) profiles.

Customers will increasingly deploy modular computing infrastructures optimized for optimum processing, memory, I/O throughput, and storage performance. Analytic capacity will be delivered in partial and full rack increments and/or via a hosted cloud offering.

BIG DATA AND YOUR DATACENTER NETWORK — ARE YOU READY?

The volume and the variety of data being generated today are growing so rapidly that CIOs are realizing they need faster, higher-quality networks not only to stay ahead of this growth but also to leverage that data for insight and business advantage. IDC believes that the network infrastructure must further Big Data's goals of high-velocity capture, discovery, and/or analysis. The best approach to accomplish this is to flatten and to implement a unified fabric architecture to meet the requirements for today's Big Data distributed processing workloads. Having storage, data management, and network resources working in concert not only brings benefits in terms of agility and scalability but also maximizes capital investments by enabling organizations to begin small pilot projects and scale cost-effectively.

Automation

No longer is data simply being stored and forgotten; rather, data of all types is being used to make proactive business decisions on a daily basis. Big Data analytics will help businesses develop more precise and timely insights, which, in turn, becomes a key business differentiator. Supporting this paradigm requires not only a fundamental shift in the way data is stored and managed by the organization but also powerful real-time data analytic and visualization tools, collaboration platforms, and automated links into existing applications that run the business, such as ERP, CRM, and financial systems.

One area where Big Data will have a direct impact on enterprise networks is in the area of network intelligence. The ability to automatically reconfigure the network for changing network loads and failed links (requiring zero administration for the addition of switch infrastructure) makes the network more agile. The result is a significantly reduced administration burden on the operations team, which minimizes the risk of errors and improves resiliency.

High Performance

The complexity of high-velocity capture, discovery, and analysis requires that Big Data architects look beyond port speed specifications and into the switch architectures individually. To ensure that the network delivers on high performance, Big Data architects test network throughput between storage, adapter, and switch. Performance is necessary for data ingestion, whether it is in bulk, micro-batching, or streaming.

Many Big Data projects also try to pull data in from transactional systems. These are real-time revenue-generating systems, and Big Data analysis has only a limited window, typically when the business is closed, to get this done. The process of extracting, transforming, and reloading data needs to happen quickly. If a project hinders the ability of a transactional system to generate revenue, it will simply fail to be supported by the organization. Network throughput is paramount to the successful implementation of Big Data in transactional environments.

Additionally, organizations must analyze the level of network intelligence offered, which furthers throughput and high performance. Questions to ask include: *Can the switch fill the pipe for bulk transfers*? *Will the architectures provide port trunking to balance the traffic load between ports, and how efficient is the use of those trunks*? *How resilient is the fabric in the event of a failure*?

Scalability

As stated earlier in the paper, most organizations are starting with smaller pilot projects and taking what they learn to fuel larger, more mission-critical efforts. As such, they need an architecture that can scale from the very small to the very large. It is imperative that they choose the appropriate network from the start. Investing in an initial small fabric will enable the network to fluidly grow with CPU and storage growth. This modular approach brings simplification to the customers and reduces operational costs.

BROCADE'S NETWORK SOLUTIONS FOR BIG DATA

The Brocade Network is designed to address holistic datacenter requirements. By delivering fabric architectures with distributed intelligence, Brocade is working to deliver a networking infrastructure that addresses the needs of today's Big Data environments — particularly for storage and compute. It provides the flexibility required to allow organizations to make the transition from Big Data as a "junior science project" to Big Data as a core business resource.

Brocade is a leader in storage networking and one of the first vendors to recognize the new requirements of Big Data environments. In the world of large content creators, Brocade's 10GbE products provide a high-performance scalable link for scale-out storage solutions such as EMC's Isilon. In the world of Big Data analytics, Brocade provides a high-performance fabric for compute solutions such as EMC Greenplum's Unified Analytics Platform (UAP).

A solution to Big Data analytics is Ethernet fabric with distributed intelligence that integrates seamlessly with existing Fibre Channel–based storage networking and Ethernet networks.

Brocade Ethernet Fabric relies on Brocade Virtual Cluster Switching (VCS) technology. Brocade's goal for Ethernet Fabric was first and foremost to provide application availability and resiliency. Additionally, the company wanted to provide a fabric that reduces infrastructure complexity while providing agility to respond to changing business requirements. Brocade VCS offers the following attributes, which are well suited for Big Data:

- Ethernet Fabric with deterministic multipathing and very low latency, without the need for complex and inefficient protocols such as Spanning Tree.
- Distributed intelligence with automatic port profile migration to seamlessly adjust Big Data configurations in the datacenter or another datacenter.
- High performance. In addition to leading industry introductions of high port speeds, Brocade offers network intelligence such as ISL Link Aggregation for customers to realize the best performance from aggregated links.
- Scalability. Brocade configurations can start with as few as 16 10 Gigabit Ethernet ports and grow up to hundreds of 10 Gigabit Ethernet ports.
- A logical chassis, which appears to network administrators as one large switch with a virtual management plane. The chassis provides increased traffic visibility and better control of the network, reducing network management overhead. While all switches in an Ethernet fabric are managed as if they were a single logical chassis, the fabric looks no different than any other single Layer 2 switch to the rest of the fabric. Each physical switch in the fabric is managed as if it were a port module in a chassis. This enables fabric scalability without manual configuration; when an additional physical switch connects to the Ethernet fabric, it automatically becomes part of the fabric and no manual configuration is necessary for setting up interswitch links.
- Dynamic services. The ability to insert services such as encryption, security, load balancing, or fabric extension while reconfiguring the network dynamically without rewiring or downtime.

Brocade technology is well-suited for Big Data platforms such as EMC Greenplum's Unified Analytics Platform. The UAP was designed to permit organizations to expedite the management of large data volumes and perform advanced analytics on Big Data sources such as structured and unstructured text, video, rich media, sensor data, and others.

Moving data inside large systems or from one system to another becomes more difficult with larger volumes. Brocade supports parallel data movement at very low latency, a key component of EMC's Big Data analytics stack. Enabled by the network, the customer's ability to get at the data and rapidly turn it into something useful for the business becomes a significant competitive advantage.

Challenges

Because organizations will want to optimize their investments holistically in the datacenter, they will ultimately seek a network that adequately supports Big Data in concert with existing virtualization and cloud deployments. As increasingly sophisticated Big Data use cases emerge with scale as the foundation, enterprises will be looking for ways to enhance their Ethernet and storage networks to meet the new scalability and automation requirements. The biggest challenge will be for Brocade to impress upon buyers that Ethernet fabric seamlessly integrates into traditional Fibre Channel and Ethernet switches and does not require a full rip-and-replace scenario. It is important that organizations understand that they can take a holistic approach to network deployments that builds upon existing best practices of storage networking while addressing the next-generation requirements of Big Data.

Conclusion

The datacenter network is in the midst of an evolution from a fixed, data-centric, client/server topology to an application-driven, dynamic network better suited to the demands of Big Data and the need for nonstop networking. By incorporating greater flexibility into the network, datacenters can improve their agility and better respond to business needs and support changing business requirements. As organizations look to maximize the opportunity for real-time analytics, they need to make sure the network is moving massive amounts of diverse data quickly and efficiently. Network performance is imperative in every aspect of the solution, including high-velocity capture, discovery, and analysis.

IDC believes that the trend toward converging network fabrics on Ethernet is the right technology path that can provide the network intelligence, ultra-low latency, and high performance needed to meet the demands of Big Data deployments.