

'Big Data' and Cloud Performance

October 11, 2010

Version 1.0

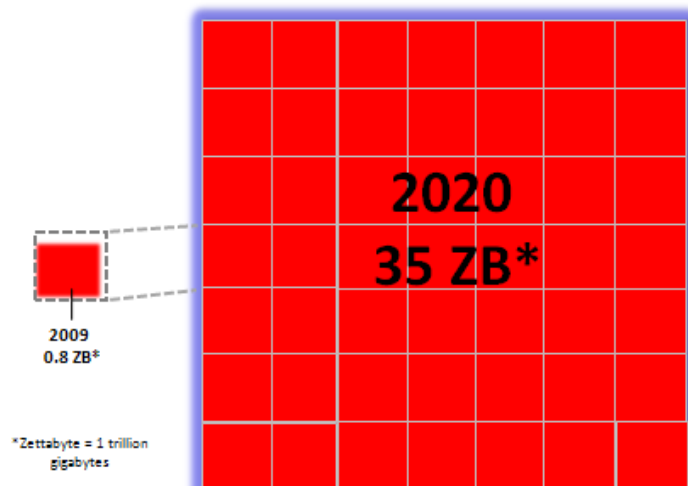
By: Russell Logan

The coming data tsunami (aka 'Big Data') threatens to swamp enterprises that are ill-prepared to manage and analyze massive data sets. Indeed even terms used to describe the quantity of data – gigabytes and terabytes, are rapidly giving way to petabyte, exabytes, and zettabytes (1 ZB = 10^{21} bytes). Petaflop/exaflop computing processing speeds and storage technologies are enabling companies to explore new business models such as in Life Sciences with genomics and personalized drug dosing. A multitude of information technologies are increasingly used by CIOs in efforts to stay afloat in a sea of data. For example, Enterprise Data Warehouse solutions (EDW) have become major growth businesses for firms like Teradata. In the past year, cloud services and distributed computing technologies offer enterprises an ability to leverage scalable compute and storage resources on a virtualized basis to help address this issue. With cloud-based storage and compute services, the notion of owning IT infrastructure (i.e., servers and data centers) may seem quaint if not somewhat bizarre in coming years. Yet with all the promise of cloud computing, looming large as a potential barrier are performance inadequacies of typical enterprise networks and public internet connections. These networks lack sufficient bandwidth to fully unleash the true potential of cloud computing technology.

What is the “data tsunami” and how big is it? A 2010 survey by IDC and EMC Corp. indicates that in 2009 consumers and businesses generated 800,000 petabytes (.8 zettabytes) of digital information. To archive this quantity of data the study suggests that it would require a stack of DVDs reaching from the Earth to the moon and back. The same study forecasts digital data creation will grow by 50% to 1.2 million petabytes (1.2 zettabytes) by the end of 2010. However by 2020, a 44-fold increase in data creation is expected and the number explodes to 35 trillion gigabytes of data — much of that in the “cloud.” That same stack of required DVDs would stretch from Earth to halfway to Mars in comparison.

Figure 1: The Digital Universe 2009 – 2020

Growing by a Factor of 44



Source: IDC Digital Universe Study, sponsored by EMC, May 2010

New technologies are at the forefront of the coming data tsunami. Technologies ranging from high fidelity sensor networks to NASA's new generation of telescopes, expected to generate an Exabyte of

data a day — definitely straining enterprise networks. Stephen Brobst, the CTO of Teradata says that a sensor data network from a Boeing jet generates 10 terabytes of information per engine every 30 minutes of flight — so for a typical cross-country flight the total amount of data generated would be a massive 240 terabytes of data. With nearly 30,000 commercial flights a day the amount of sensor data from commercial flights quickly climbs into the petabyte scale — for a single day. The obvious multiplication impacts of time results in an incredible volume of sensor data from this single application.ⁱ According to Australia's Commonwealth Scientific and Industrial Research Organization — the country's national science agency, in the next decade, astronomers expect to be processing 10 petabytes of data every hour from the Square Kilometer Array (SKA) telescope. The SKA telescope is expected to generate nearly 7 - 10 exabytes every month of operation. According to IBM, the new SKA telescope initiative will generate 3 to 4 times as more data at 1 exabyte per day. IBM is designing special hardware to process this information.

While this massive amount of data certainly presents challenges for CIOs to simply figure out storage solutions the greater issue lies with how to manage, move and analyze the data to extract meaningful information and insights, which enable companies to be more competitive. Remote storage will play a key role in managing and archiving this data. Data analytic tools and applications will be vital for enterprises to extract information and insight. While other tools provide an ability to add structure to unstructured data -- to be able to search and discover critical information (e.g. find a face in a security video). Yet without better networks search, analytics, and storage solutions in the cloud are impractical.

For enterprises dealing with these massive amounts of data network latency can be a major problem in moving the data or using cloud-computing resources. The issues related to network latency are exacerbated by geographically dispersed organizations. With globalizing businesses and 21st business models with remote and mobile workforces, providing ready access to data and applications to this decentralized workforce is daunting. Geography is a critical factor in distributing the applications and getting closer to the edge — closer to the user.

Applications that are most susceptible to latency are those, which depend most heavily on high transaction rate processes, which drive CPU per second cycles, memory and storage read/write requests, and server requests. Examples of latency sensitive applications range from multimedia streaming, video transcoding, multi-player network gaming to telesurgery and computerized trading. Other "emerging" latency sensitive applications can be seen in the high performance distributed computing space in areas such as climate weather and ocean modeling and the search for new energy and efficient solar cells.

Companies such as CFN Services, with expertise with Financial applications; among the most latency sensitive applications widely in use. They see three Tiers of latency sensitivity applications for the enterprise:

Tier I (today mostly suitable to proximity services since these are the most latency sensitive)

- Financial applications
- Primary/transactional storage
- Real-time medical imaging
- Real-time analytics
- High performance computing

Tier II (Moderate to high latency sensitivity)

- Virtual desktop
- Distributed, networked storage and compute (using resources distributed across datacenters)
- Transactional applications
- Private cloud
- Hybrid cloud

Tier III (Least latency sensitive)

- Enterprise applications
- Network backups
- Public cloud
- Collaboration
- VoIP

Multinational enterprises operating private clouds, as well as large XAAS cloud service providers, need a flexible and robust network architecture that reduces latency to acceptable levels for their particular application. WAN optimization is one approach to improve network performance and reduce application latency and conserve bandwidth. Certainly techniques such as caching compression, optimization on the protocol stack, and other various software technologies can produce incremental improvements.

Ultimately, for real performance improvements, upgrades in backbone networks are vital to create the speeds necessary to connect at multi-gigabyte-per-second speeds (or faster). A data “autobahn” linking data centers, Ethernet exchanges, and cloud services providers, to enterprises at the metro level are the best approach to handle the data tsunami. Using this approach creates logical aggregation and hubbing points where dedicated high performance connections between major cloud providers and enterprises can be established. Following this approach, CFN has created a purpose-built low latency network to facilitate high frequency trading. Using lessons learned in that demanding environment can be translated more broadly to other applications where distributed networks with low latency are required.

A multistep approach is critical to create an enterprise-specific or cloud service provider-specific network platform. The key steps include:

- Establish service delivery thresholds and metrics for services
- Optimize your service radius based on costs and end-user experience
- Position Service Nodes in proper radius of end-users
- Prioritize service scaling & go-to-market needs

For planning purposes a general rule of thumb is if the application performance works at 150ms radius, 2 or 3 locations globally will be sufficient. Separating compute and storage needs a deeper edge with roughly 2 Asia, Europe, North America, and Sao Paulo, to achieve 150ms for application performance.

ⁱ <http://gigaom.com/2010/05/04/we-cant-squeeze-the-data-tsunami-through-tiny-pipes/>