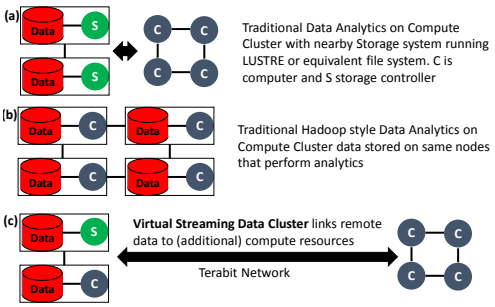
Big Data Grand Challenge for Terabit Networks

Ilya Baldin (ibaldin@renci.org, RENCI/ UNC-CH); José Fortes (fortes@ufl.edu, Florida University); Ian Foster (foster@cs.uchicago.edu, University of Chicago); Geoffrey Fox (gcf@indiana.edu, Indiana University),

Geoffrey Fox Details *http://www.infomall.org gcf@indiana.edu*
Fox received a Ph.D. in Theoretical Physics from Cambridge University and is now distinguished professor of Informatics and Computing, and Physics at Indiana University. He has supervised the PhD of 65 students and published around 1000 papers in physics and computer science with a hindex of 67 and over 23000 citations. He currently works in applying computer science to Bioinformatics, Sensor Clouds, Earthquake and Ice-sheet Science, and Particle Physics. He is principal investigator of FutureGrid – a testbed for clouds, grids and high performance computing. He is involved in several projects to enhance the capabilities of Minority Serving Institutions. He is a Fellow of APS and ACM. The description below describes the areas relevant to workshop that we are interested in. We do not request travel.

Introduction

Big Data is a dominant theme both academically and commercially. Notable recent events include a National Academy study [1], a Big Data program at NIST [2] and several other federal initiatives including a Big Data to Knowledge program at NIH [3]. Remarkable statistics [4] include over 500 million images uploaded each day in 2013, the cost of gene sequencing decreasing a factor of 1000 more than cost of computing since 2007 [5] and almost 8 zettabytes (8 million petabytes) of digital information to be stored and shared by 2015 [6]. The latter number can be contrasted with Cisco's estimate [7] that total IP traffic in 2015 will "only" be 1 zettabyte per year in 2015. Total information stored is growing faster than Moore's law while IP traffic is increasing but surprisingly slower than Moore's law. This highlights a grand challenge of Big Data emphasized at a recent NITRD MAGIC meeting [8] by Jim Pepin of Clemson [9]. Current network architectures do not allow the needed computing to be brought to this growing data deluge in an effective fashion. This computing is essential to enable the transformations (data analytics) that transform data to information to knowledge and then wisdom, policy and decisions. In this white paper, we suggest addressing this challenge using terabit networks together with software defined computing systems (virtual clusters) as shown in fig. 1(c). The software defined systems combine the growing number of dynamic provisioning tools for clusters and clouds with software defined networks. The traditional data analytics architectures in fig. 1(a) and 1(b) can only be used when data is copied to a common repository; as discussed below this is often unrealistic.



Software Defined Systems

Recently several tools have emerged to dynamically build computing environments at the level of individual nodes. These environments can be based on Bare Metal or Virtual machines and the software build includes IaaS (Infrastructure as a Service e.g. operating system) or that plus PaaS (Platform as a Service e.g. MPI and data libraries) and SaaS (Software as a Service e.g. full data analytics). These tools include templated image libraries [10], authentication and authorization, accounting and metrics [11], user interfaces (dashboards like Nimbus Phantom or OpenStack Horizon), DevOps (e.g. Salt, Chef, Puppet), dynamic provisioning [11] and higher level tools at scheduling level and above. These can be combined with tools like Rocks to manage clusters and support software defined (virtual) clusters. Further at the NaaS (Network as a Service) layer [12], technologies like OpenFlow and projects like GENI support software defined networks that allow higher performance interconnections; combining these ideas, we find software defined systems. We consider here the case of a software defined system that includes localized computing (e.g. a cluster, cloud) linked using software defined networks to distributed data.

Big Data Grand Challenge

Consider the architecture of a data repository which often in the past focused on storage and access to the data. However many researchers now need systems that manage both the (big) data and provide computing (data analytics) on the data. Here we are often told to bring the computing to the data to avoid overheads of data transport. Indeed recent commercial systems such as those at Google,

Facebook, Amazon, and Twitter are architected as clouds supporting analytics with co-located storage and satisfy the principle of bringing computing to the data. However, this is not universally applicable for data-intensive science, which features a diversity of distributed data analyzed by a distributed research team. The growing NIST Use Case collection [13] is a good source of academic, government and commercial challenges in this area and other resources are [14] and [15]. We can find examples in the LHC data analysis system with 15 petabytes of data per year distributed and analyzed across the world soon after it is first collected in a single underground cavern near Geneva, Switzerland (the LHC experiment CMS has 3,600 people, representing 183 scientific institutes and 38 countries). The primary data analysis of LHC uses compute-data affinity in a grid to produce event parameters, but also needs to access data across the globe, making it more efficient (as only parts of distributed files are needed) than transferring files. However, this model is not obviously sufficient in other disciplines where one traditionally uses repositories like GENBank (Biology), NSIDC (Polar data) and EOSDIS (Earth Satellite data), which do not always have enough attached computing to support science analysis. We can imagine adding a cluster (cloud) in front of each discipline repository but how do we determine the right size and scalability for it? How can it be elastic if it is single use? Further suppose we want to do environmental genomics; we need genomics as well as environmental data, which are typically not in same repository as they are gathered by multidisciplinary studies. One solution would be to locate all data next to the same giant compute environment – Amazon, Azure or Google or an exascale supercomputer. One might expect this to be used in some fields but it does not seem to be a general solution. It's hard to get all data co-located.

Note that in genomics, data is now gathered by a multitude of distributed low cost "individual" sequencers such as the Illumina MiSeq which has an instrument cost of ~$100K and can produce many gigabytes of data a day where distributed data to be presented to the analytics as a virtually co-located data system perhaps set up as virtual Hadoop file system running BLAST.

Disaster Recovery, HPC and Real-time Management Challenges

Another challenge that would benefit from terabit networks is that of moving large software systems from one location (e.g. datacenter) to another (due to disaster, security breaches, etc). This challenge requires large amounts of bits (describing virtual machines, data, management information, etc.) to be transferred in short periods of time, particularly in cases where one wants to keep services running. If, in addition, one needs to transfer (big) data in these moves, the bandwidth requirements will be daunting even if Terabit connections are available, unless systems are designed differently. This is also the case for distributed systems that try to emulate centralized systems that use high-speed networks possible within racks or local clusters. For example, MPI-based systems deployed on WANs fail to deliver acceptable performance whenever significant communication requirements are present; systems that use remote memory access are unrealistic in a WAN context, and delay-sensitive communication protocols do not function well in WANs. While speed-of-light limitations bound achievable latencies due to physical links, other additional latencies are due to bandwidth limitations along and at the ends of the links. These additional latencies will have to be addressed by terabit networks.

From the perspective of end nodes, significant improvements to the I/O subsystem will be needed: currently, Intel's fastest QPI performance is about 256Gbps and at least a 4x improvement is needed. Faster memory devices will be also needed - current DRAM technology has latencies on the order of 10ns, while cache memory has latencies on the order of 1ns. The notion of switching/routing and store-and-forward model of networks - i.e., the need for packet inspections - calls for network devices with extremely fast processors and memory. In terms of OS/application interface to the network, RDMA would make sense (instead of socket's send/receive model, and avoid unnecessary buffers). One could consider CPUs with network interfaces integrated into the core, alongside memory controllers. This would require redesign of MMUs and how OS manages memory (considering remote memory space). In this scenario, process or VM migration could be accomplished by simply changing the memory maps of source and destination machines. Regarding reliable lossless data transfer, can a terabit optical infrastructure offer lossless/congestion free network? If so, UDP-like low overhead protocols can be developed. Regarding low latency access to network, applications will need low latency communication mechanisms - i.e., avoid multiple bufferings and OS overheads. Applications should have access to network much like they have access to memory. Advanced RDMA mechanisms from optical infrastructure can make applications access remote data simply by issuing load/store instructions. Regarding better signaling/synchronization, CPU interrupts and/or polling are high latency mechanisms for synchronization. If a terabit optical infrastructure could offer advanced synchronization/notification mechanisms, significant advances on distributed/parallel

applications can be expected (e.g., improvements to MPI barrier and allgather). When considering parallel streams, bulk data transfers can rarely fully utilize 10Gb bandwidth. In many cases, parallel streams are used to utilize as much as possible the available bandwidth. In terabit networks, improved control of streams and help from the infrastructure will be needed (e.g., monitoring the network for feedback-based control).

Project

We suggest support of a suite of experiments that compare this streaming virtualized data model of fig. 1(c) with the traditional models of fig. 1(a,b) where files are copied from distant to local storage. One should consider multiple data architectures including databases, virtualized data stores (virtual disk images), HDFS (data parallel storage like Hbase), Lustre (wide area file systems) and the object stores exemplified by Amazon S3 and OpenStack Swift. The analytics needed include some pleasing parallel as in sensor data analysis (Cisco [16] predicts that The Internet of Things will have 50 billion devices by 2020; these are typically naturally networked to a cloud); some with weak coupling as in Genomic sequence comparison as with BLAST; others with interactive browsing as with Geographic Information Systems in Earth Science and closely coupled clustering seen in Genomics and Network Science. These analytics span performance issues including I/O dominated and communication intensive applications. Other follow-on projects could include software defined networks supporting (cloud) bursting to cope with temporary overloads, and investigation of the impact of terabit networks on cloud infrastructures (how terabit performance is exposed to virtualized resources; fast migration of a large, number of large virtual machines and storage; send/receive vs. load/store networking; and real-time constraints)

Acknowledgements

These ideas have been developed with Mauricio Tsugawa, Gregor von Laszewski, and Martin Swany

References

1. Committee on the Analysis of Massive Data; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; National Research Council, *Frontiers in Massive Data Analysis*. 2013: National Academies Press. http://www.nap.edu/catalog.php?record_id=18374
2. NIST. *Big Data Initiative*. 2013 Available from: http://bigdatawg.nist.gov/home.php.
3. NIH. *Big Data to Knowledge (BD2K) initiative*. 2013 Available from: http://bd2k.nih.gov/.
4. Mary Meeker and Liang Wu (Kleiner Perkins Caufield Byers), *Internet Trends*, in *D11 Conference (All Things Digital)*. May 29, 2013. Rancho Palos Verdes, California. http://www.slideshare.net/kleinerperkins/kpcb-internet-trends-2013.
5. NIH. *DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program.* [accessed 2013 August 14]; Available from: http://www.genome.gov/sequencingcosts/.
6. IDC iView (sponsored by EMC). *Extracting Value from Chaos*. 2011 June [accessed 2013 August 14]; Available from: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.
7. Cisco. *Visual Networking Index: Forecast and Methodology, 2012–2017*. 2013 May 29 [accessed 2013 August 14]; Available from: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html.
8. NITRD. *MAGIC Meetings 2013*. 2013 [accessed 2013 August 14]; Available from: http://www.nitrd.gov/nitrdgroups/index.php?title=MAGIC_Meetings_2013.
9. Geoffrey Fox, Miron Livny, Shantenu Jha, Dennis Gannon, Ioan Raicu, and Jim Pepin. *Minutes of Presentations on 2020-2025 Scientific Computing Environments (Distributed Computing in an Exascale era)*. 2013 August 7 [accessed 2013 August 14]; Available from: http://www.nitrd.gov/nitrdgroups/images/7/7a/Geoffrey_Fox_MAGICMinutesAugust72013.pdf.
10. Javier Diaz, Gregor von Laszewski, Fugang Wang, and Geoffrey Fox, *Abstract Image Management and Universal Image Registration for Cloud and HPC Infrastructures*, in *IEEE CLOUD 2012 5th International Conference on Cloud Computing* June 24-29 2012. Hyatt Regency Waikiki Resort and Spa, Honolulu, Hawaii, USA http://grids.ucs.indiana.edu/ptliupages/publications/jdiaz-IEEECloud2012_id-4656.pdf
11. Gregor von Laszewski, Hyungro Lee, Javier Diaz, Fugang Wang, Koji Tanaka, Shubhada Karavinkoppa, Geoffrey C. Fox, and Tom Furlani, *Design of an Accounting and Metric-based Cloud-shifting and Cloud-seeding framework for Federated Clouds and Bare-metal Environments*, in *Workshop on Cloud Services, Federation, and the 8th Open Cirrus Summit*. September 21, 2012. San Jose, CA (USA). http://grids.ucs.indiana.edu/ptliupages/publications/p25-vonLaszewski.pdf.
12. E. Kissel and M. Swany, *Evaluating High Performance Data Transfer with RDMA-Based Protocols in Wide-Area Networks*, in *14th International Conference on High Performance Computing and Communications (HPCC-2012)*. June 25-27, 2012. Liverpool UK. http://damsl.cs.indiana.edu/projects/phoebus/sdn_xsp.pdf.
13. NIST. *Collection of Big Data Use Cases (still being collected)*. 2013 August 14 [accessed 2013 August 14]; Available from: http://bigdatawg.nist.gov/_uploadfiles/M0105_v3_5876831460.docx.
14. Geoffrey Fox, Tony Hey, and Anne Trefethen, *Where does all the data come from?* , Chapter in *Data Intensive Science* Terence Critchlow and Kerstin Kleese Van Dam, Editors. 2011. http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%20all%20the%20data%20come%20from%20v7.pdf.
15. Geoffrey Fox. *Big Data X-Informatics MOOC*. 2013 August 14 [accessed 2013 August 14]; Available from: https://x-informatics.appspot.com/course.
16. Cisco Internet Business Solutions Group (IBSG) (Dave Evans). *The Internet of Things: How the Next Evolution of the Internet Is Changing Everything*. 2011 April [accessed 2013 August 14]; Available from: http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf.