Open Problems in Network-aware Data Management in Exa-scale Computing and Terabit Networking Era

Mehmet Balman
Lawrence Berkeley National Laboratory
Berkeley, CA, USA
mbalman@lbl.gov

Surendra Byna
Lawrence Berkeley National Laboratory
Berkeley, CA, USA
sbyna@lbl.gov

ABSTRACT

Accessing and managing large amounts of data is a great challenge in collaborative computing environments where resources and users are geographically distributed. Recent advances in network technology led to next-generation high-performance networks, allowing high-bandwidth connectivity. Efficient use of the network infrastructure is necessary in order to address the increasing data and compute requirements of large-scale applications. We discuss several open problems, evaluate emerging trends, and articulate our perspectives in network-aware data management.

Categories and Subject Descriptors

C.2.4 [**Computer-Communication Networks**]: Distributed Systems—*Distributed applications*

General Terms

Design, Performance

Keywords

network-aware tools, resource provisioning, high-bandwidth networks, data-intensive, distributed computing

1. INTRODUCTION

Communication is a necessity for the human race; hence, with the course of time, the network, the Internet, and eventually the high-speed networks had been invented. But as the rate of generated electronic data rose exponentially over the years, concomitant with the increase in storage capacity and network bandwidth, so did the network usage and traffic. This brought in a whole new set of challenges to be dealt with, before high-speed high-bandwidth terabit networks and exa-scale computing systems can become a reality. Some of the challenges include:

- Can we use the same protocols and algorithms in terabit networks that work for everyday Internet?
- What stumbling challenges do we expect in exa-scale data management on terabit networks?
- Do we need to re-engineer existing tools and distributed middleware systems in order to adapt exa-scale data management on terabit networks?
- How are we going to handle network management in terms of provisioning capacity and path, performance monitoring and diagnosis in the future?

NDM'11, November 14, 2011, Seattle, Washington, USA.

In this high-speed network age, network-aware data management (NDM) is emerging as a prominent field, especially with the soaring volume of scientific data pushed daily into the network from important applications like climate modeling, astronomy simulations, bio-computation, and high-energy physics research.

Efficient data access is crucial, both in operating system design and in microprocessor architecture. In operating systems, we move pages from disk to memory; in microprocessor architecture, instruction fetch is important for performance; in supercomputers, optimizing I/O accesses has a major performance impact; on large-scale distributed systems, transferring data between geographically separated storage sites has great importance on the overall end-to-end system performance.

Data management has been one of the crucial problems in every stage of computer engineering, from micro level to macro level. Today, we face the same problem, but at a much larger scale, and likewise, we need to reconsider our traditional approaches to deal with handling data in next-generation networks.

In this paper, we discuss several open problems, evaluate emerging trends, and articulate our perspectives in usage of networking for distributed data management. In doing so, we reflect back to the paper contributions and panel discussions at the 1^{st} Network-aware Data Management Workshop (NDM 2011) held in conjunction with the IEEE/ACM International Conference for High Performance Computing, Networking, Storage and Analysis (SC 2011), where we got a glimpse of the state-of-the-art. We classify the challenges in network-aware data management into the following items.

- Increasing amounts of scientific data
- Need for efficient use of high-bandwidth networks
- Network provisioning and traffic isolation
- Novel data access mechanisms and models
- Ease-of-use and user needs

In the following sections, we briefly discuss each of the challenges and solution trends.

2. A QUICK OVERVIEW OF DATA-DRIVEN SCIENCE

Data-intensive science spans over multiple disciplines including climate, biology, astronomy, and high-energy physics. In general, we can briefly explain the life-cycle of scientific data in three main steps. The first step is the collection of data either from an experimental facility or a scientific analysis device such as genome sequencers, or a scientific application such as climate or astronomy simulations. Due to the rapidly advancing capabilities of computing and sophisticated hardware, these data generators produce enormous amounts of data. The next step is transferring and storing the data temporarily or permanently for further processing and scientific analysis. Depending on where the analysis of the data takes place, data transfers play a quintessential role. It is a common practice that analysis of data occurs on computers away from the origins of data, requiring large data transfers over network. In the final step, the produced analysis-output is archived, and in most cases, shared in a community storage for reproduction and verification of the results.

In a nutshell, current science requires large data, large computation, and also large collaborations among multiple research groups to produce high-impact discoveries and scientific solutions.

2.1 Exponential growth of scientific data

Nowadays, domain scientists are generating terabytes (TB), and even petabytes (PB) of data every month. As mentioned by Brian Tierney in the NDM 2011 [1] panel, data traffic is expected to reach 100 petabytes/month in Energy Sciences Network (ESnet) by 2015. The scientific process consists of several components coupled into a single, complex application that may need to adapt dynamically to changing environmental conditions. For example, fusion research requires executing a complex workflow management system including several varieties of analysis applications. Each component generates output files that serve as input to other analysis applications. One of the main challenges is the increasing data size in such a coupled simulation environment. In metagenomics, next-generation sequencing devices are generating data more than the capacity of compute power. Devices are distributed in separate research facilities where multiple users analyze data with multiple software tools.

In climate research, several HPC centers generate simulation data and publish the data over several gateways to make the data available to the community worldwide. Researches download a particular subset of data of their interest into their local site and process the data locally to evaluate results. The demand for accessing and distributing scientific data is also increasing exponentially. For example, more than 25,000 users downloaded data from the Intergovernmental Panel on Climate Change (IPCC) Forth Assessment Report [2]. IPCC data is hosted among multiple sites over the world, and downloaded and shared by many users worldwide. The recent Coupled Model Intercomparison Project, Phase 5 (CMIP-5) is estimated to be 1.5-2PB. Due to the increased demand for climate research, more data downloads are expected in the future.

The increasing data raises the importance of network-aware data management. We need advanced middleware tools for managing distributed compute power and storage capacities. Distributed computing environment has many benefits including reliable data access, load balancing, access to many compute resources, and use of multiple data centers. Mostly, multiple geographically distributed institutions are involved for the scientific data processing, and data need to be shared and distributed over the network. Recent technology provides high-bandwidth interconnects, but efficient use of high-performance networks is crucial for end-to-end overall application performance. As mentioned by Daniel S. Katz in the NDM 2011 panel [1],'Network will never be fast enough'. The number of data intensive applications is increasing and the amount of data is growing in rate faster than the available bandwidth capacity. Hence, network-aware data management is gaining importance by the years, as the scientific collaboration and data sharing, and as a result, network usage increases.

3. OPEN CHALLENGES IN NDM

3.1 How to enable high-performance data movement?

Efficient utilization of network resources is an essential requirement of network-aware data management. High-speed optical networks are reaching 100Gbps capacity. However, large-bandwidth provided by today's networks cannot be fully utilized due to inadequate protocols, poorly tuned protocol parameters, and underutilized capacity in end-systems. In a distributed workflow for a large-scale application, inability to adequately use the data communication might result in delays in other participating components. This results in slowing down the entire system and making it inefficient. Thus, there have been many efforts to optimize network transfers with application level tuning [3, 4], to keep the network pipe full for maximum throughput. One common approach is to use parallel streams, where multiple threads work simultaneously to overcome the latency cost. Another approach is to use concurrent transfers in which multiple transfer nodes cooperate together to generate high throughput data in order to fill the network pipe.

Dynamic data transfer optimization is a challenging and highly desired feature among the network-aware data management tools. In most cases, the transfer node is the bottleneck [5]; slow performance in I/O operations through the local file system, memory latency and duplicate buffer copy operations result in performance degradation in end-to-end data transfer. A common challenge is finding the optimum parameter to tune and achieve high throughput with minimum load. As mentioned by Ann Chervenak in the NDM 2011 panel [1], the parameters such as parallel streams, buffer sizes, and protocol and kernel tuning need to be adjusted according to the capacity of the underlying environment. The achievable end-to-end throughput and the system load in communicating parties might change during the period of a data transfers, especially when large volume of data needs to be transmitted. Therefore, dynamic approaches in which data transfer tuning is performed on the fly [6], are highly desirable in order to adapt to varying environmental conditions to come up with a high-quality tuning for best system and network utilization.

In addition to multiple streams and parameter tuning, request-aggregation is important. Many small data movement requests are combined and embedded into a single operation to increase overall performance, especially for trans-

fers of data sets with overwhelming number of small files. Without this optimization, each transfer operation requires initializing the network transfer protocol and setting up a connection to a remote data transfer service. When it adds up, this combination of initialization and connection setup time becomes a significant portion in the total cost. As the number of requests for small amounts of data increases, aggregating them into a large transfer utilizes the bandwidth of the high-speed networks more efficiently.

3.2 Is RDMA a feasible option in wide-area?

Remote direct memory access (RDMA) and iWARP are promising technologies that have benefits applicable to next-generation networks. In iWarp, the TCP stack is incorporated within hardware (i.e. TCP offload engine) to reduce the burden on local CPU. RDMA enabled network adapters directly copy into/from remote memory. The main idea in zero-copy networking and RDMA is to minimize the interaction with CPU, caches, etc. and to reduce the network protocol overhead [7]. The transfer operations can be done asynchronously, so that the transfer latency is reduced, thus enabling fast messaging [8]. One of the main motivations is to minimize the effect of memory latency and to eliminate the involvement of host CPU for the transfer operations. Nowadays, RDMA over Converged Ethernet (RoCE) is gaining high interest for efficient data transfer with low latency over loss-less Ethernet networks. Among other things, technology trends, benefits, opportunities, and usage of RDMA in high-speed networks was discussed by Dhabaleswar Panda in the NDM 2011 panel [1].

RDMA has been heavily applied in local area networks especially to enhance the communication between compute and storage nodes in a cluster. Local area networks and wide area networks have different characteristics, so they demonstrate diverse features in terms of congestion, failure rate, and latency. RDMA provides direct end-to-end one-sided data movement between application memories. The current experiments with RDMA over wide-area, require a dedicated virtual circuit isolated from other traffic. We still require better understanding and integration with QoS network reservation systems in order to evaluate it as an alternative protocol for high-performance data transfer in wide-area. In addition to that, most of the bulk data transfer applications work in a streaming fashion. Therefore, it is quite important to present a network-aware efficient cache management and aggregation logic in order to utilize the underlying message-oriented RDMA technology.

3.3 Latency: Achilles' heel of high-bandwidth networks

Although there is a tremendous effort in implementing high-bandwidth networks such as 40Gbps and 100Gbps interconnects, a challenging issue is to use the underlying network infrastructure in efficient manner. The performance of an end-to-end data transfer operation not only depends on the available bandwidth, but it is also very much related to the latency and performance of the participating transfer nodes. Even though we are able to add high memory and cache capacity in today's systems, latency still remains as one of the challenges in terms of performance. For example, in microprocessor, in order to improve execution throughput, we hide memory latency by pipelining. Analogous to the memory latency, we are limited with the network latency in widely distributed systems.

Underlying transfer protocols in today's system are not ideal for transferring large amounts of data in a shared wide-area network over long distance connections. Using multiple data transfer streams is a popular technique by users to increase the total transfer throughput. Instead of a single connection at a time, multiple streams are opened for a single data transfer service. Larger bandwidth is gained but concurrent data transfer operations that are initiated at the same time over-provision the network and system resources. Especially in a commodity network where bandwidth is shared, this results in poor performance for other users. Current technology enables high-bandwidth wide-area connections, but latency between interconnects is one of the bottlenecks in high performance data access. What further actions do we need to take in order to bring true high-performance networks into a reality? As the network bandwidth increases and more scientists/users involve in collaborative work, the effect of latency will be felt deeper.

3.4 Predictable Performance vs. Guaranteed Bandwidth: Which one?

High bandwidth networks are one of the major components needed to enable large-scale data replication, high performance remote data analysis and visualization, and providing access to computational resources. In next-generation terabit network infrastructure, dedicated communication channel will most-probably play an important role by providing guaranteed bandwidth and traffic isolation for efficient resource utilization. There had been several projects developed to provide high-speed on-demand data access between collaborating institutions. For instance, the ESnet's bandwidth reservation system, OSCARS [9], establishes a dedicated virtual circuit with guaranteed bandwidth between two end-points over a certain amount of time. User requests are received over a web interface and a QoS secure virtual path is created using Multi-Protocol Label Switching (MPLS) and the Resource Reservation Protocol.

The fair sharing of the network in a best-effort approach possibly causes ineffective use of the available network bandwidth. Delivering network-as-a-service that provides communication capabilities in a secure, reliable, and dedicated manner has many benefits. A promising approach is to isolate the bulk data transfer from best-effort traffic. In guaranteed bandwidth, we can use the on-demand dynamic circuit setup between participating hosts for high-bandwidth data access. This enables full utilization of the underlying network resources and also coordinates and tunes participating host system for end-to-end performance.

Better user interfaces, and novel methodologies are required for flexible reservations and end-system integration such as provisioning of transfer nodes and storage capacities. Flexible reservation is analogous to the scenario in reserving airline flights such that flight reservation systems offer number of options based on the possible date/time requirements of the traveler. In the past, we proposed a new methodology [10] in which network provisioning system can make optimal selections based on user-provided time and resource constraints, such as the earliest time in which data will be ready, the deadline for completion of the task, the expected size of total data transfer. Also, in the NDM 2011 workshop, multiple papers explained the current develop-

ment of network-aware tools and new services for network provisioning to support high-performance data movement.

From users' point of view, a system that performs 'well-enough' is sufficient in most cases. Instead of guaranteed bandwidth and isolated traffic, predictable performance may suit better in which a user plans ahead, and chooses the time-interval that the data movement operations can be accomplished in a reasonable manner, without breaking the application constraints. The importance and benefits of predictable behavior have been emphasized by Richard Carlson in the NDM 2011 panel [1]. We need analytical models to interpret application behavior and network usage in multi-domain systems. Considering the fact that QoS systems are not widely deployed, network performance modeling and performance prediction need to be studied to the optimize use of networking, and also to identify major system bottlenecks. Do we need complex QoS control mechanisms for network provisioning and bandwidth reservation, or predictable behavior with reasonable performance is all scientists need?

3.5 Envisioning next-generation data-access models

Research efforts in new data access mechanisms are necessary. Most scientific data is distributed among many researchers, and the high-bandwidth networks will only be limited to the connections between several institutions in the initial phases. Hence, research should also be focused on enhancing data distribution in the Internet and improving data download rates. Efficient data replication techniques should be developed to place data on separate data nodes that are accessed in a federated manner. In addition to using redundant resources to improve reliability, we also need to replicate data for application performance in order to have faster data access. However, replication comes with its storage cost and requires a management system for coordination and synchronization. It is beneficial to bring data easily on-demand when needed such that data is retrieved into the application in real-time. Remote I/O over wide-area may suffer from poor latency. By keeping the network pipe full by feeding enough data into the network, hiding the effect of network latency and improving the performance is possible. Advanced data staging and data access techniques would play an important role in order to use remote storage resources efficiently over the network. Intelligent data flow coordination and data caching that are designed according to the application behavior, are significant in the use of networking for high-performance data management.

As computation hardware performance grew at a rapid pace over the last few decades, storage hardware (mainly hard disk drives) performance improvements have been very slow. In parallel systems, accessing data from storage costs tens of thousands of CPU cycles and the corresponding energy wastage in waiting for data to arrive for processing. If data access has to become available at real-time in distributed environments, the network latency adds up to the storage access cost. The common practice is to move all the files needed for analysis regardless of what fraction of data is useful from the transferred files. For example, our recent study envision to move data from remote storage through querying mechanisms [11]. Using such a data service, applications can directly access remote data by specifying attributes of the required data without transferring files.

3.6 Understanding user needs

Next-generation, high-bandwidth networks bring new challenges such as heterogeneous resource layers, parallelization of capacity, managing aggregation, and utilizing multiple links over intra-domain boundaries. The system software should take advantage of the underlying parallel file systems, network reservations for the middleware, and WAN-transfer tuning. Advanced knowledge is required in many cases to tune applications for network usage and configure system components for high-performance data access. Hence, understanding user needs and building an overall user-friendly system are stumbling challenges.

The tools developed for network-aware data management must be easy-to-use for achieving larger user base. During the NDM 2011 keynote speech, Ian Foster emphasized that traditional techniques of leaving the burden on the user for moving/storing data is not a viable option anymore. For example, the Globus Online service [12] lets users drag-and-drop files between remote locations in an online interface. The fire-and-forget style data transfer feature of this service attracted many users recently. Simplifying end-to-end data flow by providing transparent easy-to-use network services is necessary. Overall, the future systems need to shift the burden of optimization and management of the network from users to the system.

4. SUMMARY

There is a need for efficient use of the network infrastructure in order to address the increasing data and compute requirements of large-scale applications. Advanced services for optimization and tuning large-scale data movement, end-to-end resource coordination and network provisioning, and novel abstraction techniques for data representation are essential in order to support the requirements of data-intensive applications these days. Since the amount of data is continuously growing, traditional techniques to manage data between distributed resources are not sufficient. Many scientific applications do poorly and fail to provide adequate use of the available bandwidth in an end-to-end context. Large-scale applications necessitate intelligent data-flow management to deal with the diverse performance requirements of different resources involved in end-to-end data movement. The lack of network-aware advanced tools in the scientific community causes ineffective use of the network. Network-aware services for managing end-to-end data flow between distributed resources are necessary in order to deliver true exa-scale performance to the application layer. All these issues will feed the next-generation research in the burgeoning field of network-aware data management (NDM).

Acknowledgments

We would like to thank Ian Foster (Argonne National Laboratory / University of Chicago) for providing his inspiring keynote speech; Richard Carlson (DoE Office of Advanced Scientific Computing Research), Ann Chervenak (USC Information Sciences Institute), Daniel S. Katz (University of Chicago and Argonne National Laboratory), Dhabaleswar Panda (Ohio State University), and Brian Tierney (ESnet / Lawrence Berkeley National Laboratory) for their presentations in the panel discussion. We thank Arie Shoshani (Lawrence Berkeley National Laboratory) for his valuable guidance in organizing the NDM 2011 workshop. We also

thank the NDM 2011 authors for their contributions, and the workshop audience for their insightful comments and feedback.

Disclaimer

The views and opinions expressed herein belong to the authors of this paper and do not necessarily state or reflect those of the NDM 2011 presenters, the keynote speaker, the panelists, or any organization, or any agency.

5. REFERENCES

[1] International Workshop on Network-Aware Data Management, in conjunction with the ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC 2011), Seattle, WA, Nov 14th, 2011. http://sdm.lbl.gov/ndm/2011

[2] Williams et al., Earth System Grid Federation: Infrastructure to Support Climate Science Analysis as an International Collaboration, to appear in Data Intensive Science Series: Chapman & Hall/CRC Computational Science, 2012.

[3] T. Yoshino et al., Performance Optimization of TCP/IP over 10 gigabit Ethernet by Precise Instrumentation. In Proceedings of the ACM/IEEE conference on Supercomputing, 2008.

[4] Ito, T.; Ohsaki, H.; Imase, M.;,On Parameter Tuning of Data Transfer Protocol GridFTP for Wide-area Grid Computing, 2nd International Conference on Broadband Networks, vol., no., pp.1338-1344 Vol. 2, 7-7 Oct. 2005.

[5] W. Allcock et al. The Globus Striped GridFTP Framework and Server, In Proceedings of the ACM/IEEE conference on Supercomputing, 2005.

[6] Balman, M.; Kosar, T.; ,Dynamic Adaptation of Parallelism Level in Data Transfer Scheduling, Complex, Intelligent and Software Intensive Systems, 2009. CISIS '09. International Conference on , vol., no., pp.872-877, 16-19 March 2009.

[7] Neeser, F. D.; Metzler, B.; Frey, P. W.; ,SoftRDMA: Implementing iWARP over TCP kernel sockets, IBM Journal of Research and Development , vol.54, no.1, pp.5:1-5:16, 2010.

[8] P. Lai, H. Subramoni, S. Narravula, A. Mamidala and D. K. Panda, Designing Efficient FTP Mechanisms for High Performance Data-Transfer over InfiniBand, Intl Conference on Parallel Processing (ICPP '09), Sept. 2009.

[9] Chin P. Guok, David Robertson, Mary Thompson, Jason Lee, Brian Tierney, William Johnston, Intra and Interdomain Circuit Provisioning Using the OSCARS Reservation System, Third International Conference on Broadband Communications, Networks, and Systems, IEEE/ICST, 2006.

[10] Balman, M.; Chaniotakisy, E.; Shoshani, A.; Sim, A.; , A Flexible Reservation Algorithm for Advance Network Provisioning,High Performance Computing, Networking, Storage and Analysis (SC), 2010 International Conference for , vol., no., pp.1-11, 13-19 Nov. 2010.

[11] K. Wu, S. Byna, D. Rotem, and A. Shoshani, Scientific Data Services - A High-Performance I/O System with Array Semantics, 1st Workshop on High-Performance Computing meets Databases (HPCDB 2011), Co-located with Supercomputing 2011, Seattle WA.

[12] Foster, I.; , Globus Online: Accelerating and Democratizing Science through Cloud-Based Services, Internet Computing, IEEE , vol.15, no.3, pp.70-73, May-June 2011