Analysis and selection of a network service for a scientific data distribution project

Jie Li, Malathi Veeraraghavan, and Matthew Manley
University of Virginia
Charlottesville, USA

Steve Emmerson
University Corporation for Atmospheric Research (UCAR)
Boulder, USA

Abstract—The volume of scientific data collected by instruments, from experimental studies, and from simulations executed on high-performance computing platforms is growing rapidly. Scientists need to move these data files from the sources where they are generated to their laboratory compute clusters for further analysis. Such scientific data transfers are “heavy-hitter” flows that consume an unfairly large portion of network resources, and thus adversely affect general-purpose flows on IP-routed networks. The development and deployment of new types of network services offer additional options for these data transfers. The problem is to develop a methodology for identifying the most suitable type of network service for each such scientific project. This paper presents a case study for this problem, in which we describe our analysis of the real-time meteorology data distributed by UCAR in the Internet Data Distribution (IDD) project, and select a suitable network service for this project. Based on this analysis, our conclusion is to experiment with reliable multicast virtual circuit service for the IDD application.

Index Terms—scientific data distribution; dynamic virtual circuit service; multicast

I. INTRODUCTION

In recent years, the size of scientific data collected by instruments, through large-scale experiments, such as the Large Hadron Collider (LHC) project [1], and from massive simulations run on super-computers, has been growing rapidly. This data is distributed to geographically dispersed scientists from the experimental/super-computing sites primarily across research-and-education IP-routed networks, such as Internet2 [2], ESNet [3], CERNET [4], etc. This traffic is often characterized as consisting of “heavy-hitter” flows that can potentially have negative effects on general-purpose IP-routed flows.

Several research-and-education network (REN) providers have deployed high-speed optical circuit/virtual-circuit networks, and offer dynamic circuit services as a complement to their IP-routed service [5]. Some scientific projects have leased static circuits from these providers to interconnect their key scientific data generation-consumption sites. The REN providers anticipate offloading these high-rate, large-sized scientific data flows to circuits, thus mitigating any adverse effects these flows may have on general-purpose flows if they are carried on their IP-routed networks.

In this work, we consider the question of identifying the best network service for scientific data distribution. Specifically, we present a case study, in which we analyze real-time meteorology data distributed by the University Corporation for Atmospheric Research (UCAR) in its Internet Data Distribution (IDD) project [6], and select a suitable network service for IDD. The IDD application is a near real-time data distribution system to over 160 institutions. The software used for this data distribution is called the Local Data Manager (LDM) [7]. Currently over 30 types of scientific data products (called feedtypes in IDD terminology) are distributed using LDM.

Our findings are as follows. Some of the feedtypes are such that new data products are created (e.g., from radar measurements) almost continuously. Static circuits from the upstream LDM server that sends the data products to the downstream LDM servers that receive the data seems to be the best matched network service. However, other feedtypes are such that data products appear in bursts, i.e., in some minutes the volume of data generated is orders of magnitude higher than in other minutes. User requirements are for near real-time delivery. Furthermore, data products need to be disseminated from one upstream LDM server to 10s to 100s of receivers. High-speed circuits are required in those minutes when the data rate is high, but for other minutes, if the same circuits are used, the utilization will be poor. We recommend the use of multicast virtual circuits rather than unicast static or dynamic circuits for this particular data distribution to save both computing and bandwidth resources relative to today’s usage of unicast TCP connections over IP-routed paths. The methodology developed here can be used to answer this question for other types of data distribution.

Section II provides the reader background material on different types of network services. Section III presents data analysis results for two representative feedtypes distributed by IDD. Section IV discusses the selection of a network service best suited for this data distribution project. Section V concludes this work.

II. BACKGROUND

In this section, we briefly review different types of network services. The first service is the ubiquitous IP-routed service, which when coupled with TCP offers users reliable data delivery. The second is the leased-line/virtual private network service, also called static circuit service, which offers users a dedicated circuit between two or more specific endpoints, for a pre-specified duration. The third type of network service, which was recently introduced by commercial and REN providers, is a dynamic circuit service (DCS). DCS provides
on-demand, rate-guaranteed communication service. To access DCS, a customer needs to first purchase a DCS access link through which the customer’s endpoint (typically an IP router) is connected to the service provider’s circuit/virtual circuit switch. Once this access link is in place, the customer can request virtual circuits to any other DCS-connected endpoints for rate-guaranteed communication on a short-term basis.

Network services can be unicast from a single destination or multicast from one source to multiple destinations. IP-routed, static circuit, and dynamic circuit services can all be configured to operate in unicast mode or multicast mode. While the term “circuit” is used for the second and third types of services, more generally, virtual circuit technologies such as MultiProtocol Label Switched (MPLS) are used for these services.

Finally, instead of unicast client-server TCP transfers, a peer-to-peer (P2P) model can be used in which all LDM servers participate as peers. After an initial push of different data products to different receivers, all receivers could become senders and send their received products to other peers. This would lower the total latency required for data distribution, and as demonstrated in [8], the latency becomes almost independent of the number of receivers.

III. DATA ANALYSIS

A high-level review of the 30 feedtypes served by IDD from the real-time statistics site [9] shows that the CONDUIT and NEXRAD2 feedtypes are two representative “heavy-hitter” flows in terms of data volume and duration. Therefore, this section provides a detailed analysis of the data transfer patterns of the CONDUIT and NEXRAD2 feedtypes.

A. CONDUIT Feedtype Analysis

Single Flow Characteristics: To characterize the data transfer patterns of a single CONDUIT flow, we installed the LDM software (version 6.8.1) on one of our laboratory machines to receive the CONDUIT feedtype from one of the data servers located at the UCAR site (idd.unidata.ucar.edu). We ran LDM in verbose mode so that information about every single data product received was recorded in a local log file. The information collected in the log files was then parsed to analyze the characteristics of the CONDUIT data flow. Based on an analysis of 9 days of data (October 13, 2010-October 21, 2010), the results showed that the CONDUIT data flow has a relatively static daily transfer pattern. Fig. 1 shows a typical transfer pattern of a single CONDUIT data flow in one day. First, the total size of CONDUIT data generated and distributed to all receivers every day is about 60 GB, with a standard deviation (SD) of 0.3 GB. Second, the throughput varies significantly at different times of the day. The maximum throughput is about 250 MB per minute (an average of 33.3 Mbps) with an SD of 28.8 MB, while the minimum throughput is 0 MB per minute. Furthermore, there are four periods during a day, separated approximately by 6 hours, when throughput is at its peak level. Third, the number of data products generated per minute is large although most data products are small in size. The per-day average number of data products is 1.6 million (SD: 0.01 million), with up to 8000 data products (SD: 1054) transferred per minute during peak flow periods. The per-day average size of all data products is about 38 KB (SD:0.2 KB).

Data Distribution Topology: The next step is to analyze the data feed topology, or Feed Trees in IDD terminology, which shows the sender-receiver relationships. The hierarchical feed tree topology for each feedtype can be obtained from the IDD real-time statistics Web site [9]. We downloaded and parsed the real-time topology data for CONDUIT. The results are shown in Table I. The real-time feed tree is a dynamic topology in that hosts join and leave the tree. The results in this table were parsed from the feed tree information obtained on April 15, 2011. The total number of distinct hosts in the CONDUIT feed tree was 163. Of these distinct hosts, 57 were sending data to other hosts (by running LDM upstream processes), while 141 were receiving data from other hosts (by running LDM downstream processes). Hosts at the intermediate levels of the feed trees are both data senders and receivers. Of all the data senders, the one located at UCAR has a maximum fan-out1 of 104. Since LDM creates a separate TCP connection for each sender-receiver pair, the peak bandwidth requirement for UCAR’s access link is 104 × 250 MB/minute, or about 3.5 Gbps just for the CONDUIT data distribution. Fig. 2 shows a visualized topology map which we downloaded from the IDD real-time statistics Web site. The central point shown in Colorado (with the maximum fanout) is the IDD site at UCAR.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of Distinct Hosts</td>
<td>163</td>
</tr>
<tr>
<td>Number of Sender Hosts</td>
<td>57</td>
</tr>
<tr>
<td>Number of Receiver Hosts</td>
<td>141</td>
</tr>
<tr>
<td>Maximum Fan-out Number</td>
<td>104</td>
</tr>
</tbody>
</table>

1Fan-out is the number of data receivers connected to the same sender.
Silence Periods: The results show that only a small percentage (less than 2%) of the time intervals between the transfer of two consecutive data products are larger than 1 second. Per day, there are only about 100 silence periods that last longer than 60 seconds out of a total of 1.6 million silence periods. Of these 100, about 15 are larger than 300 seconds. The maximum silence period observed across all 9 days of traffic was 35 minutes.

B. NEXRAD2 Feedtype Analysis

Single Flow Characteristics: We collected information about the NEXRAD2 feedtype using the same methodology described for the CONDUIT feedtype. Data was collected for the period Sept. 15, 2011 - Sept. 21, 2011. As with the CONDUIT feedtype, the data generation and transfer patterns of NEXRAD2 shows little variation from one day to the next. Fig. 3 shows the per-minute throughput and number of products for Day 7 (Sept. 21, 2011). The per-day average data size is 56.80 GB, with a standard deviation of 6.56 GB. Unlike CONDUIT, the variability of the NEXRAD2 throughput is relatively small. The per-minute average throughput varies from 4.0 Mbps to 7.8 Mbps.

Data Distribution Topology: The data distribution topology for the NEXRAD2 feedtype on a typical day is shown in Table II. The total number of hosts involved in NEXRAD2 data distribution is 150, with a maximum fanout of 55, which again corresponds to the LDM servers located at UCAR. Given the observed maximum throughput of 7.8 Mbps, this creates a peak bandwidth requirement of $55 \times 7.8$ Mbps, or about 429 Mbps, for the NEXRAD2 feedtype.

Silence Periods: There was only 1 silence period per day that had a duration longer than 300 seconds. This period occurred at the end of the day and lasted 5250 seconds. There were a total of about 81240 silence periods per day. Almost all of these silence periods were less than 1 second.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of Distinct Hosts</td>
<td>150</td>
</tr>
<tr>
<td>Number of Sender Hosts</td>
<td>75</td>
</tr>
<tr>
<td>Number of Receiver Hosts</td>
<td>114</td>
</tr>
<tr>
<td>Maximum Fan-out Number</td>
<td>55</td>
</tr>
</tbody>
</table>

TABLE II

NEXRAD2 FEED TREE TOPOLOGY INFORMATION

IV. SELECTION OF A SUITABLE NETWORK SERVICE

Currently, IDD uses unicast TCP/IP service from an upstream LDM server to all its downstream LDM servers. For example, as shown in Table I, for the CONDUIT feedtype the UCAR LDM servers send data to 104 downstream LDM servers. It does this via unicast TCP connections to each downstream server. Multiple processes are forked from the main upstream server, which means effectively, data products are served in round-robin fashion to all 104 receivers. The use of this service type has the key advantage of leveraging an already deployed and widely used network service. However, it requires UCAR to run 9 hosts for IDD, and uses 5 Gbps of the UCAR access link. Furthermore, the latency in delivering the data products is sensitive to the number of receivers. Finally, as a relatively heavy user of the UCAR access link, this traffic likely has adverse effects on general-purpose IP-routed flows. Therefore, we consider the other network service options reviewed in Section II.

First, consider the use of static unicast virtual circuits. As shown in Fig. 3, NEXRAD2 data products are generated almost continually, making static virtual circuit service a good match. However, as described in Section I, CONDUIT data is bursty, and has silence periods. If the circuit rates are chosen to lower latency for the large burst minutes, the circuit utilization will be poor in the other minutes. Furthermore, just for CONDUIT and NEXRAD2 data, as described in Section III, UCAR needs at least 4 Gbps on the access link.

Next, consider the use of dynamic circuit service. Since the...
data patterns are quite predictable, scheduled dynamic circuit service could be used for CONDUIT data. However silence periods, as described in Section III are on the order of seconds, while circuit setup delay in today’s deployment on RENs such as Internet2, and ESnet, is on the order of minutes. If data products are buffered and sent on high-speed circuits, for the worst-case fanout of 104 for CONDUIT, the total delay will be greater than with IP service, since for each receiver a new circuit needs to be set up.

While the use of circuit services will offload these scientific heavy-hitter flows from the IP-routed network, thus reducing the adverse effects of these flows on general-purpose flows, they still incur the other disadvantages of unicast IP services: latency dependence on the number of receivers, the need for multiple hosts and high access link bandwidth at senders such as UCAR. Therefore, we consider multicast services. With multicast service, both access-link bandwidth and computing resource requirements will be lower. The question, however, is whether it is feasible to design a multicast solution in which one slow receiver does not decrease the throughput for all receivers. A new Virtual Circuit Multicast Transport Protocol (VCMTP) has been designed and prototyped with this feature of slow receivers not reducing the throughput of fast receivers [10].

Finally, when comparing multicast virtual circuit service with P2P service, we anticipate lower delays with the former. The latter offers an advantage over unicast client-server model, but it nevertheless requires more than one transfer for most of the blocks in a data product, while the former requires one transfer for most blocks with retransmissions due to bit errors or receive buffer overflows for a few blocks. P2P service is suitable for files that are obtained by different peers at different times, but with IDD, as soon as a new data product is created, all subscribers require near real-time delivery, for which multicast virtual circuits seem better suited.

V. CONCLUSION

The selection of a suitable network service for a specific scientific data distribution task depends on the data characteristics, data distribution topology, and performance requirements of the task. In a case study, we analyzed the characteristics of two representative feedtypes in the Internet Data Distribution (IDD) Project, which is used to distribute meteorology data to over 160 institutions from UCAR, and argued that multicast virtual circuits are most appropriate for this application. Work is ongoing to develop a transport protocol for multicast virtual circuits, and future work is planned to integrate this service with the Local Data Manager (LDM) software used in the IDD project.

ACKNOWLEDGMENT

The University of Virginia portion of this work was supported by the NSF grants OCI-1038058 and OCI-1127340, and DOE grants DE-SC002350 and DE-SC0007341.

REFERENCES