The Design and Implementation of an Available Bandwidth Measurement Scheme in the K*Grid System

Seong-il Hahm*, Seongho Cho*, Han Choi*, Chong-kwon Kim*, and Pillwoo Lee

Abstract: Grid computing is an emerging technology that enables global resource sharing. In Korea, the K*Grid provides an extremely powerful research environment to both industries and academia. As part of the K*Grid project, we have constructed, together with the Korea Institute of Science and Technology Information and a number of domestic universities, a supercomputer Grid test bed which connects several types of supercomputers based on the globus toolkit. To achieve efficient networking in this Grid testbed, we propose a novel method of available bandwidth measurement, called Decoupled Capacity measurement with Initial Gap (DCIG), using packet trains. DCIG can improve the network efficiency by selecting the best path among several candidates. Simulation results show that DCIG outperforms previous work in terms of accuracy and the required measurement time. We also define a new XML schema for DCIG request/response based on the schema defined by the Global Grid Forum (GGF) Network Measurement Working Group (NM-WG).

Keywords: K*Grid, Globus Toolkit, Available Bandwidth Measurement, XML schema

1. Introduction

High performance computing technology is of great importance in modern industry and science. For instance, scientists seek new drugs for cancer or AIDS with the aid of supercomputers. However, ever with a 70 TFlops supercomputer, which is the most powerful supercomputer in the world according to top500.org, scientists still need a faster computer because there are always more complex problems which cannot be solved by the most powerful computer in the world. Grid computing offers a solution to this problem. Grid computing is a form of distributed computing that connects computers, data storages, and other scientific facilities across geographically dispersed organizations, enabling huge computing resources, vast data repositories, and virtual experimental laboratories.

All over the world, many Grid-related projects have been carried out. For example, GriPhyN [6] provides a computational resource for major scientific experiments in physics, astronomy, biology and engineering in the U.S., Europe, and Asia. NASA's Information Power Grid (IPG) [9] is a high-performance computation and data Grid that integrates geographically distributed computers, databases, and instruments. In Korea, the K*Grid [11] provides an extremely powerful research environment to both industry and academia. The K*Grid project has developed a national Grid infrastructure based on supercomputers and high performance clusters. As part of this development, we have constructed a supercomputer-based Grid test bed by connecting several types of supercomputers. Heterogeneous supercomputers are logically incorporated into a Grid system by operating the globus toolkit[‡]. The test bed successfully tested a local supercomputer's ability to command remote ones to execute a script written by the local machine in Resource Specification Language (RSL).

With the supercomputer test bed, we studied how network performance could be optimized for file transferring. To resolve this issue, we proposed a new method for estimating the end-to-end available bandwidth of the network. This method decouples the effects of network topologies or competing traffics from the delay patterns of the packet trains which passed along the bottleneck link. We then obtained the bottleneck capacity and the amount of competing traffics, from which we were able to calculate the available bandwidth as the bottleneck capacity minus the amount of competing traffics. We named this method the Decoupling Capacity measurement with the Initial Gap (DCIG). Also, we proposed a Fast Converging (FC) algorithm to measure the available bandwidth more rapidly. Considering the linearity of the input gap of a packet pair with their output gap, we found the input gap which reflects the competing traffic. From this technique, we can reduce the number of packet trains to measure the available bandwidth.

The Network Measurement Working Group (NM-WG) [3] of the Global Grid Forum (GGF) [1] has developed the

Manuscript received October 4, 2005; accepted December 20, 2005.
This work was partially supported by the Brain Korea 21 Project in 2006. The ICT at Seoul National University provided research facilities for this study.

Corresponding Author: Seong-il Hahm
* School of Electrical Engineering and Computer Science, Seoul National University, Seoul, Korea ({siham, shcho, hchoe, ckim}@popeye.snu.ac.kr)
** Dept. of Supercomputing, Korea Institute of Science and Technology Information (KISTI), Daejeon, Korea (pwlee@kisti.re.kr)

[‡] The Globus Toolkit was developed by Globus Alliance for building Grid systems and applications [4].

XML schema for publishing Grid-network measurement data. This NM-WG schema defines the request/response formats for network metrics or such characteristics as bandwidth capacity, available bandwidth, one way delay, etc. Based on the NM-WG schema, this paper also defines a new schema (dcig.xsd) for DCIG request/response.

The rest of the paper is structured as follows. Section 2 outlines the K*Grid project and its supercomputer-based test bed. In Section 3, we propose a new method for measuring end-to-end available bandwidth. Based on the NM-WG schema, we define a new schema for DCIG request/response in Section 4. Section 5 describes other Grid projects and work related to end-to-end available bandwidth measurements. The conclusions are given in Section 6.

2. K*Grid and its Test Bed

The K*Grid project [11] is an initiative in Grid research supported by the Korean Ministry of Information and Communication. The main goal of the K*Grid project is to provide an extremely powerful research environment including a huge amount of computing power, virtual experiment facilities, and international collaborations, to both industry and academia. The K*Grid project includes Grid applications, middleware, and the development of the national Grid infrastructure.

2.1 Middleware and Applications of K*Grid

MoreDream is a Grid middleware which integrates many geographically and organizationally dispersed computing resources, massive data capacity, and human resources for the K*Grid. This middleware uses globus toolkits as a reference model to efficiently integrate the developed components and to verify the functions.

Grid applications are essential to test the state-of-the-art Grid technologies, to measure the performance of the K*Grid infrastructure, and to conduct new requirements for further development. Then, with the help of the advanced K*Gird infrastructure, innovative scientific research programs are expected. Many Grid applications, such as nano material computing and bio-technologies, are studied in the K*Grid project to efficiently fuse Grid technologies and application.

2.2 Grid Infrastructure

The national Grid infrastructure consists of several types of supercomputers such as Massively Parallel Processor (MPP) machines and cluster-based computers which are dispersed throughout domestic universities and national institutes. As part of the test bed, Seoul National University connected its MPP-based machines (IBM SP Nighthawk-II) to a linux cluster-based one at the Korea Institute of Science and Technology Information. These heterogeneous supercomputers are logically incorporated into a Grid system by operating the globus toolkit 2.4.2 [4] on each supercomputer.

While the globus toolkit consists of open source software, the architecture and operating system of the IBM machines are not open. This mismatch caused several problems in installing the globus toolkit on the IBM machines. The most severe problem was that the operating system of the IBM machines (AIX 4.3.3), which had already been installed at the beginning of the construction of the test bed, did not support the globus toolkit 2.4.2. Note that this globus version is the target of the K*Grid because it is a stabilized version. To surmount these problems, we upgraded the operating system to AIX 5.1 before installing globus. After setting up the globus toolkit, MPICH-G2 1.2.5 was also installed to support remote job assignments. Our test bed successfully tested the ability of a local supercomputer to command remote ones to execute a script written by the local machine in Resource Specification Language (RSL), and received the corresponding results from the remote machines.

3. A New Available Bandwidth Measurement Technique

In the K*Grid test bed, we studied how the network performance could be optimized when file transferring occurs from one side to another. To achieve efficient networking, available bandwidth is the most important information for Grid applications. To measure the available bandwidth of a network path, both centralized and distributed methods exist. Well-known centralized measurement tools include the Multi Router Traffic Grapher (MRTG) [14] based on Simple Network Management Protocol (SNMP) [15]. These methods can obtain accurate information directly from routers. However, this privilege is limited only to the network administrator in a single Autonomous System (AS). Therefore, a multiple Internet Service Provider (ISP) environment or distributed end-to-end applications cannot use centralized measurement information. Also, there are several distributed measurement techniques. These distributed measurement techniques are classified into bottleneck capacity measurement, available bandwidth measurement, and bottleneck link placement. These mechanisms purely use the end-to-end probing information obtained by sending packer pairs or packet trains.

Currently, several available bandwidth measurement methods have been considered in the Global Grid Forum (GGF) and its eXtensible Markup Language (XML) schemas have been defined. However, given that the available bandwidth techniques take too much time to measure an available bandwidth, we propose a new available bandwidth measurement mechanism by considering the exact bottleneck capacity and the fast convergence of the measurement algorithm. To measure the available bandwidth accurately, we must remove the effect of network topologies or tight link/narrow link while measuring the bottleneck link capacity. Also, the linearity of the input gap of a packet pair with the output gap of the pair can reduce measurement time. From the above observation, we propose the Decoupled Capacity measurement with Initial Gap (DCIG) technique and the Fast Converging (FC) method.

Our proposed mechanisms are based on single-hop gap model [17]. In this model, a network path is modeled as a single bottleneck link and the gap of a packet pair is mainly affected by the bottleneck link. In this model, the output gap of packet trains can be increased by the competing traffic. If the competing packets are queued after the first probing packet arrival, the second probing packet of a packet pair experiences a queuing delay at a bottleneck router. In this case, the increased gap of a packet pair is proportional to the ratio of the competing traffic to the bottleneck capacity. This relation can be formulated as follows

$$g_o = \frac{p + B_i g_i}{C}, \tag{1}$$

where p is a probing packet size, B_i is the transmission rate of the competing traffic, and C is the capacity of the bottleneck link.

Previous methods measure the bottleneck link capacity C from the back-to-back packet trains. However, back-to-back packet trains can suffer from a distortion of the output gap by the network topologies or the competing traffics. To solve this back-to-back packet train problem, we propose a decoupled capacity measurement method with the initial gap of the packet train. To decouple the effect of network topologies or competing traffics, we obtain the bottleneck capacity C and the amount of competing traffics B_i from two samples of $<g_i, g_o>$ pairs. From Equation (1), the measured bottleneck link capacity and the amount of competing traffics is thus

$$C = \frac{p(g_i^1 - g_i^2)}{g_i^1 g_o^2 - g_i^2 g_o^1}, \quad B_i = \frac{p(g_o^1 - g_o^2)}{g_i^1 g_o^2 - g_i^2 g_o^1}, \tag{2}$$

where $<g^1{}_i, g^1{}_o>$ and $<g^2{}_i, g^2{}_o>$ are the first and second samples of packet trains respectively. Therefore, we can obtain the bottleneck capacity C and the amount of competing traffics B_i from the two samples. Also, we can obtain the available bandwidth A_i from the extraction $C - B_i$. We call this algorism the Decoupled Capacity measurement with Initial Gap (DCIG) method.

Also, the gap difference $g_d = g_o - g_i$ has a linear relationship with the initial gap g_i. We can obtain the gap difference $g_d = g_o - g_i$ relationship as follows

$$g_d = \frac{g_d^1 - g_d^2}{g_i^1 - g_i^2} g_i + K, \tag{3}$$

where $d^1{}_d$ and $d^2{}_d$ represent the output and input gap difference of the first and second sample respectively. From Equation (3), we can predict the input gap g_i which satisfies the output, and the input gap difference g_d becomes zero. From this procedure, we can reduce the probing packets. We call this mechanism the Fast Converging (FC) method.

Combining the above two methods, we can measure the available bandwidth of the network path accurately and quickly compared to self-induced congestion methods like TOPP [12], pathChirp [16], and Pathload [10]. Self-induced congestion methods transmit probing packets to the bottleneck link and observe the point of delay inflection. These mechanisms take too long to measure the fluctuating network path bandwidth [17]. Also, single-hop gap model-based mechanisms - including our proposed mechanism - can measure quickly, but can be inaccurate because of the existence of a tight/narrow link problem [10]. However, our mechanism overcomes the problem of IGI by decoupling the effect of the network topologies or the competing traffics when measuring the capacity of the bottleneck links.

We compare the performance of DCIG, IGI/PTR [7], and pathChirp [16] by computer simulation using an ns-2 simulator [13]. We use a three-hop topology link as shown in Fig. 1.

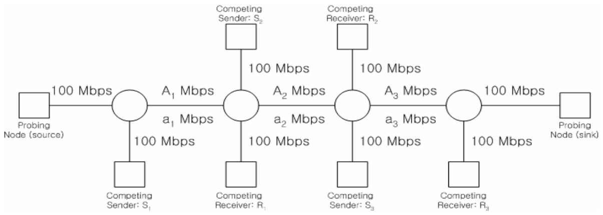

Fig. 1. Simulation Topologies

We select the bottleneck bandwidth pair, (A_1, A_2, A_3) as (40, 60, 40). The competing traffic rates a_1 and a_3 are fixed at 5 Mbps and as a_2 varies from 30 to 45 Mbps. From Fig. 2, we can find the DCIG method to measure the available bandwidth faster and more accurately. Because pathChirp measures the available bandwidth by congesting the bottleneck link, we can observe that pathChirp converges slowly to obtain the available bandwidth result, and less accurately compared to single-hop gap model based methods like IGI and DCIG. Also, the DCIG method can measure more accurately compared to the IGI method.

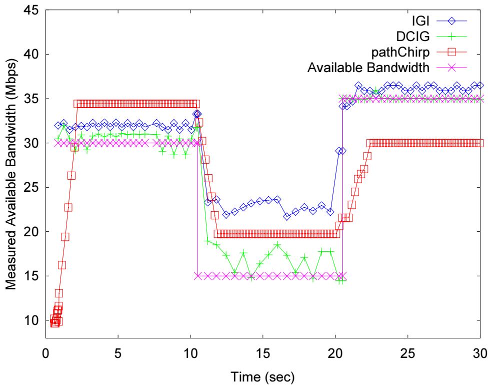

Fig. 2. Convergence and Accuracy of Measurement Methods

4. Schema for DCIG Request/Response

The Network Measurement Working Group (NM-WG) [3] of the Global Grid Forum (GGF) [1] has developed the XML schema for publishing Grid-network measurement data. This NM-WG schema defines the request/response formats for network metrics or characteristics such as bandwidth capacity, available bandwidth, one-way delay, etc. Based on the NM-WG schema, we define a new schema (dcig.xsd) for DCIG request/response as shown in Fig. 3. When a client wants to know the available bandwidth between any two nodes, it sends a request XML file to a server including "subject" element which has a "src" and "dst" child element, and a "parameters" element which has both the "packetsizeintrain" and "trainlength" child element. Both the "packetsizeintrain" and "trainlength" elements are allowed max and min values. A response XML file contains the results of the "path. bandwidth.available" element, which is defined by NM-WG, and a "timerequired" element, which means the time required to measure the available bandwidth. Due to the limited space, we omit the examples of the request/response XML files.

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema
xmlns:xs="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified"
targetNamespace="http://www.ggf.org/nmwg/tools/DCIG/"
xmlns:nmwg="http://www.ggf.org/nmwg/"
xmlns:topology="http://www.ggf.org/nmwg/topology/"
xmlns:DCIG="http://www.ggf.org/nmwg/tools/DCIG/">
  <xs:import       namespace="http://www.ggf.org/nmwg/"
schemaLocation="nmbase.xsd"/>
  <xs:import
namespace="http://www.ggf.org/nmwg/topology/"
schemaLocation="topology.xsd"/>
  <xs:element                              name="metadata"
substitutionGroup="nmwg:Metadata">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="DCIG:subject"/>
        <xs:element ref="DCIG:parameters"/>
      </xs:sequence>
      <xs:attribute         name="id"          use="required"
type="nmwg:Identifier"/>
    </xs:complexType>
  </xs:element>
  <xs:element name="subject" type="DCIG:DCIGSubject">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="topology:src"/>
        <xs:element ref="topology:dst"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element                            name="parameters"
type="DCIG:DCIGParameters">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="packetsizeintrain">
          <xs:attribute name="size">
            <simpleType>
              <restriction base="xs:integer">
                <minInclusive value="40"/>
                <maxInclusive value="1500"/>
              </restriction>
            </simpleType>
          </xs:attribute>
        </xs:element>
        <xs:element name="trainlength">
          <xs:attribute name="length">
            <simpleType>
              <restriction base="xs:integer">
                <minInclusive value="20"/>
                <maxInclusive value="256"/>
              </restriction>
            </simpleType>
          </xs:attribute>
        </xs:element>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="datum">
    <xs:complexType>
      <xs:sequence>
        <xs:element       name="path.bandwidth.available"
type="nmwg:BandwidthAvailableType"
substitutionGroup="nmwg:resultSet"/>
        <xs:element                    name="timerequired"
type="xs:string"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

Fig. 3. Schema (dcig.xsd) for DCIG request/response based on the NM-WG schema

5. Related Work

Major Grid projects are being conducted throughout the world. First, the GriPhyN [6] project brought together a team of Information Technology (IT) researchers and experimental physicists to provide the IT advances required to enable petabyte-scale data intensive computing. The project was driven by unprecedented requirements for the geographically dispersed extraction of complex scientific information. To meet these requirements, the GriPhyN team pursued IT advances centered on the creation of Petascale Virtual Data Grids (PVDG) that could meet the data-intensive computational needs of scientific problems. Second, NASA's Information Power Grid (IPG) [9] is also representative of high-performance computing and data Grids, integrating geographically distributed computers, databases, and instruments. Like the electric power Grid for which it was named, the IPG delivers computational power on the basis of who needs the power, rather than where the power is located. The members of the IPG project at the NASA Ames and Glenn Research Centers are performing research and development to help create a robust, functional, and easy-to-use computational data Grid for NASA.

Measuring bandwidth is important for use of the Grid because bandwidth information can be used in performance monitoring services or Grid networking optimization. GGF NM-WG determines the mapping of some common network measurement tools to the network characteristics described in [2]. We explain briefly four representative measurement tools which measure the available bandwidth of a specific network path.

Iperf [8] measures the maximum TCP bandwidth, allowing the tuning of various parameters, by actually sending a large amount of TCP-emulated packets. However, these emulated packets may cause large probing overheads. Pathload [10] is based on the basic idea that one-way delays of a periodic packet stream show an increasing trend when the stream rate is larger than the available bandwidth. The measurement algorithm is iterative and it requires the cooperation of both the sender and the receiver. IGI/PTR [7] models the Internet path as a single hop and uses the packet train to predict the available bandwidth of bottleneck links. It can predict the available bandwidth faster than Pathload, but measurement results can have large errors due to capacity estimation errors. pathChirp [16] is based on the concept of "self-induced congestion," where pathChirp features an exponential flight pattern of probes called a chirp. By rapidly increasing the probing rate within each chirp, pathChirp obtains a set of information from which to dynamically estimate the available bandwidth.

6. Conclusion

As Grid computing is an emerging technology for enabling global resource sharing, international organizations operate their own Grid projects. Korea also has the K*Grid [11] project which provides an extremely powerful research environment to both industry and academia. As part of the K*Grid project, we have constructed a supercomputer Grid test bed which consists of heterogeneous supercomputers by using the globus toolkit. To achieve efficient networking in this Grid test bed, we implemented a novel method for available bandwidth measurement, called Decoupled Capacity measurement with Initial Gap (DCIG), using packet trains. Our proposed mechanism can provide fast and accurate available bandwidth measurement results compared to self-induced congestion methods like TOPP [12], pathChirp [16], and Pathload [10], since it decouples the effect of network topologies or competing traffics while measuring the capacity of bottleneck links. We also define a new XML schema for DCIG request/response based on the NM-WG schema.

Reference

[1] GGF, http://www.Gridforum.org, 2005.

[2] GGF NM-WG, “A Hierarchy of Network Performance Characteristics for Grid Applications and Services”, GWD-C, Community Practice, Jun. 2003.

[3] GGF NM-WG, http://nmwg.internet2.edu, 2005.

[4] Globus Alliance, http://www.globus.org, 2005.

[6] GriPhyN, http://www.griphyn.org/, 2005.

[7] Hu, N. and Steenkiste, P., “Evaluation and Characterization of Available Bandwidth Probing Techniques,” IEEE Journal on Selected Areas in Communications, Vol. 21. No. 6, 2003.

[8] Iperf, http://dast.nlanr.net/Projects/Iperf, 2005.

[9] IPG, http://www.ipg.nasa.gov/, 2005.

[10] Jain, M. and Dovrolis, C., “End-to-End Available Bandwidth: Measurement Methodology, Dynamics, and Relation with TCP Throughput,” ACM SIGCOMM, 2002.

[11] K*Grid, http://www.gridcenter.or.kr/, 2005.

[12] Melander, B., Bjorkman, M. and Gunningberg, P., “A New End-to-End Probing and Analysis Method for Estimating Bandwidth Bottlenecks,” IEEE Global Internet Symposium, 2000.

[13] NS-2, http://www.isi.edu/nsnam/ns, 2005.

[14] Oetiker, T. Multi Router Traffic Grapher (MRTG), http://people.ee.ethz.ch/~oetiker/ webtools/mrtg

[15] Presuhn, R., Case, J., McCloghrie, K., Rose, M. and Waldbusser, S., “Version 3 of the Protocol Operations for the Simple Network Management Protocol (SNMP),” IETF RFC 3416.

[16] Ribeiro, V., Riedi, R., Baraniuk, R., Navratil, J. and Cottrell, L., “pathChirp: Efficient Available Bandwidth Estimation for Network Paths,” Passive and Active Measurements (PAM) workshop, 2003.

[17] Strauss, J., Katabi, D., Kaashoek, F., “A Measurement Study of Available Bandwidth Estimation Tools,” ACM SIGCOMM Internet Measurement Workshop, 2003.

Seong-il Hahm
He received the BS degree in computer science and electrical engineering from Handong University and the MS degree in electrical engineering and computer science from Seoul National University in 2002 and 2004, respectively. He is currently a PhD candidate in the School of Electrical Engineering and Computer Science at Seoul National University, Seoul, Korea. His current research interests are wireless LANs, mobile ad-hoc networks, wireless mesh networks, and wireless opportunistic scheduling.

Seongho Cho

He received the BS degree in Department of Computer Sceience and the MS degree in school of Electrical Engineering and Computer Science from Seoul National University, in 1999, 2001, respectively. He is currently a PhD candidate in the school of Electrical Engineering and Computer Science at Seoul National University. His research Interests are mobility management, wireless MAC, and network measurement.

Han Choi

He received the BS degree and the MS degree in mathematical science and in electrical engineering and computer science from Seoul National University in 2003 and 2005, respectively. He is currently a PhD candidate in the School of Electrical Engineering and Computer Science at Seoul National University, Seoul, Korea. His current research interests are wireless measurements and transfer protocols in highspeed networks.

Chong-kwon Kim

He received the BS degree in industrial engineering from Seoul National University, the MS degree in operations research from Georgia Institute of Technology, and the PhD degree in computer science from the University of Illinois at Urbana-Champaign in 1981, 1982, and 1987, respectively. In 1987, he joined Bellcore as an MTS in applied research. He worked on Broadband ISDN switching, ATM QoS support, and traffic management problems in Bellcore. Since February 1991, he has been with Seoul National University as a professor in the School of Computer Science and Engineering. His current research interests are wireless networks, highspeed network control, distributed processing, and performance evaluation.

Pillwoo Lee

He received the BS degree in electrical engineering from Dongguk University and the MS degree and the PhD degree in computer engineering from Tsukuba University in 1988, 1991, and 1999, respectively. Since April 2004, he has been with Korea Institute of Science and Technology Information (KISTI) as the Grid Computing Research Team Leader in Supercomputing Center. His current research interests are computer architecture, grid computing, and distributed computing.