

Grid Operation at Tokyo Tier-2 Centre for ATLAS

Hiroyuki Matsunaga, Tadaaki Isobe, Tetsuro Mashimo, Hiroshi Sakamoto & Ikuo Ueda

International Centre for Elementary Particle Physics, the University of Tokyo, Japan

Abstract

International Centre for Elementary Particle Physics, the University of Tokyo, has been involved in the Worldwide LHC Computing Grid since 2003. After extensive R&D experience of the PC computing farm, disk and tape storage systems, network technology and the integration of these components, it is now operating a regional centre for the ATLAS data analysis. The regional centre includes an ATLAS Tier-2 site which is running the gLite middleware developed by the Enabling Grids for E-science (EGEE) project. One of the biggest challenges at the regional centre is efficient data transfer between the Tier-2 site in Tokyo and other sites, in particular the associated Tier-1 site in France, because the large round trip time due to the long distance makes it difficult to transfer data at a high rate. We have been studying to achieve a good performance of the data transfer, and some results of network tests and ATLAS data transfer are described. Hardware and software components and the operational experience are also reported in this article.

1. Introduction

International Centre for Elementary Particle Physics (ICEPP), the University of Tokyo was established originally in 1974, aiming at studying the most fundamental particles and forces of nature experimentally by using the world's most for front accelerators. In the past, ICEPP has been involved in several collider experiments at DESY and CERN in Europe, and now the main project is the ATLAS experiment at the Large Hadron Collider (LHC). The LHC is the world largest accelerator which has been constructed at CERN near Geneva and is almost ready to start operations. The first collision will be expected in 2008.

At the LHC, data volume coming from the detector becomes huge due to the high-energy and high-luminosity proton beams and a huge number of readout channels of the detectors. The amount of data produced at the LHC experiments will be 15PB per year, and the processing power will require more than 10 million Spe-

cInt 2000 (SI2000). Consequently the experiments require very large data storage capacity, very high data processing speed and very wide data transfer bandwidth. As a result, even CERN, the world largest accelerator laboratory, can't afford to provide the necessary computing resources for the data analysis. On the other hand, the LHC experiment collaborations have been very large and international, and recent development of the Wide-Area Network (WAN) connectivity enabled distributed data analysis by connecting computing centres of the collaborating institutes around the world. The LHC experiments decided to utilize the Grid technology for the distributed computing framework to analyze their data. The Worldwide LHC Computing Grid (WLCG) was created to enable the data analysis for LHC experiments with the Grid. It is an international collaboration of the physics laboratories and institutes which contribute resources, the four LHC experiments (ALICE, ATLAS, CMS, LHCb), and the Grid projects providing software (EGEE, OSG and Nordugrid).

In Japan, it was decided that ICEPP should operate a regional centre for ATLAS at the University of Tokyo to enable efficient data analysis by the Japanese collaborators. At present ICEPP is running a Tier-2 Grid site (named TOKYO-LCG2) as a member of the WLCG collaboration, and is also providing a local analysis facility for the Japanese collaborators in ATLAS. The Tier-2 site has been involved in the WLCG and ATLAS distributed computing activities for the last few years. Overview of the regional centre and the recent operational experience will be shown below.

2. *Hardware*

We, at ICEPP, have been studying on the computer fabric and the site operation by using the pilot systems since 2002[1]. In this study, PC farms, disk storages, tape drives and network devices were tested extensively, and expertise of the Grid deployment has been acquired. Based on this study the production system for the regional centre was procured by international tendering in 2006, and started running from January 2007.

The production system (Fig. 1) has been placed in the new computer room (_270m²) dedicated to the regional centre. Electric power is supplied through uninterruptible power supplies (UPS) to all computing components as well as air conditioners. Under a room space constraint, blade server is used in the PC farm (Grid worker nodes and local batch nodes) and Grid head nodes thanks to the high density and low power consumption. We introduced Dell PowerEdge 1955, with dual Xeon 3160 CPUs (WoodCrest 3.0GHz, dual core), 4 or 8GB memory (2GB/core is necessary for the ATLAS reconstruction program) and 73GB local Serial Attached SCSI (SAS) disk drives in a mirroring (RAID-1) configuration by hardware RAID. There are 650 blade nodes in total.

As for the data storage, an external RAID system (Infortrend EonStor A16FG2422) was chosen based on good performance experience in the pilot system, as a cost-effective solution in order to have as much disk space as possible. It contains 16 Serial ATA (SATA) hard disk drives (500GB each) and has a 4Gb Fibre Channel host interface. The disk storage can be accessed through a disk server (Dell PowerEdge 2950, 8GB memory) which is equipped with a 10GbE network interface card (Chelsio T210-SR) for the network connection and a 4Gb Fibre Channel host bus adapter (QLogic QLE2462). Both the RAID systems and the disk servers are connected to Fibre Channel switches (QLogic SANbox 5600). Each disk server serves 5 filesystems (35TB in total), each of which is provided from the RAID system configured as one RAID-6 partition. We have 140 RAID systems which correspond approximately to 1PB.

Sun StorageTek SL8500 is our tape library system. In the present configuration, it has 8000 tape slots and 32 tape drives, and LTO 3 tape cartridges (400GB capacity) are used. Each tape drive are connected to a tape server (Dell PowerEdge 1950) with a 2Gb Fibre Channel host bus adapter (QLogic QLA2342-CK). This system is currently setting up as a component of a hierarchical storage management (HSM) system, and will be used at the local facility in the near future.

There are four Ethernet switches (three Foundry RX-16's and one RX-4) for the data transmission in the regional centre. For connections with the blade servers, we use 1000Base-T high-density line cards with mini RJ-21 interfaces to reduce number of cables.

ICEPP signed the WLCG Memorandum of Understanding (MoU), in which we pledged to provide 1000kSI2000 of CPU, 200TB of disk and 2000Mbps of WAN connection in 2007. In 2008 additional 200TB of disk is pledged to provide and it is now in preparation. The next major hardware upgrade will be done by 2010 at ICEPP, and the pledged numbers in the MoU would change in future according to the actual ATLAS activity and the next procurement result.



Fig. 1 Picture of the computer room. The blade servers and the RAID systems are mounted in the racks. A tape library system is placed at the far side.

3. *Software*

At our regional centre, a part of resources is provided to the Tier-2 Grid site and the rest (local facility) is used by the Japanese collaborators only. These two parts are logically separated and we manage these two in different ways, except for Operating System (OS) installation and fabric monitoring. The local facility can be accessed even without the Grid software, and we will provide local users with data accesses to the data storages between the two domains.

3.1 Grid middleware

The gLite middleware developed by EGEE is used at the Tier-2 site. The gLite 3.0 has been installed on Scientific Linux CERN 3 (SLC3) OS with 32bit architecture on all nodes but Worker Nodes (WNs) and Storage Element (SE) nodes.

On the WNs, gLite 3.1 has been installed on SLC4 32bit OS since 2007 as it was requested from the ATLAS. As for the SE, the Disk Pool Manager (DPM) has been deployed since it is recommended to use at a Tier-2 site (not having a tape storage) and it is relatively easy to operate. We have one head node and six disk servers for

the SE. On the head node, several important services (srmv1, srmv2, srmv2.2, dpm, dpns and MySQL as a backend database) are running, while actual data transfer is performed on the disk server on which a GridFTP daemon is running. The disk server has gLite 3.0 installed on SLC4 x86 64 OS because the OS allows us to make a 6 TB filesystem on a RAID partition and Linux kernel 2.6 in SLC4 should be better than kernel 2.4 in SLC3. For the filesystem of the DPM disk pool, we introduced XFS (by SGI) which handles large files better and is faster in deleting a large amount of data files than ext3, which is the default file system in recent Linux distributions. Also, XFS requires 64bit OS due to its implementation. In the burn-in tests before the deployment, we found that XFS in the SLC4 was not stable enough in case a filesystem was almost full; hence we had to modify some kernel parameters to avoid a kernel crash. The DPM recently implemented the Storage Resource Manager (SRM) v2.2 protocol, and this has been configured with the space tokens at ICEPP as requested by ATLAS.

As the local batch system in Computing Element (CE), we are using Torque and Maui included in the gLite middleware. We have no major problem with 480 CPU cores in 120 WNs, but if we want to provide more CPU cores and/or WNs we may migrate to LSF (Load Sharing Facility) of Platform that is already in use at the local facility.

In addition to the above-mentioned services, we have top-level and site BDIIs, MON, RB (Resource Broker) and MyProxy, LFC (LCG File Catalogue) with MySQL backend, and UI (User Interface) nodes in production at our site.

We have a testbed for the gLite middleware deployment. The middleware upgrade sometimes requires a configuration change and the gLite middleware itself as well as its configuration tool (YAIM) have often bugs. We usually try to upgrade the middleware on the testbed before doing this on the production system in order to validate it. It is very useful to avoid instability or an unnecessary downtime.

3.2 Other softwares and tools

For the local analysis facility, we have different strategy than the Tier-2 Grid site. Since its computing resource is larger than that of the Tier-2 site and data access pattern is different, performance and functionality are sometimes more important than stability.

In the CPU farm in the local facility, we are using LSF as the batch system, although Torque/Maui was mainly used in the past pilot systems and is also used at our Tier-2 site even now. The reason is that numbers of CPU cores are so many that we have not tried with Torque/Maui, and Torque/Maui does not provide sufficient functionalities for fine-grained configuration. We have been running LSF for more than a year, and we would like to deploy it in the Tier-2 site as well in the future.

The local disk storage can be accessed by NFS (Network File System) version 3 at present since we have had good experience with it and many users like having access to a filesystem instead of using a special IO protocol. However, as data volume becomes large we want to integrate the tape system in our storage system by using CASTOR (CERN Advanced Storage Manager) version 2, a HSM system developed by CERN. We carried out some tests with CASTOR 1 with the pilot system, and are currently setting up CASTOR 2 in the production system. After the functional tests we will start using CASTOR 2 soon, and perhaps will replace DPM with CASTOR 2 at the Tier-2 site in future.

A key component of the CASTOR is the Oracle relational database. In the production system of the regional centre, we set up Oracle 10g with 2-node RAC (Real Application Clusters) configuration. It is to be used with CASTOR and the gLite middleware, and possibly for ATLAS conditions databases by replicating from another Tier-1 site. We are still getting experience of Oracle with the FTS (File Transfer Service) in gLite.

We provide 10 interactive nodes at the local facility. To allow for balancing loads, round robin DNS is used for them. They also act as gLite UI nodes and submission nodes to the LSF batch system. Users' home directories and ATLAS software are served by NFS, but we are trying to migrate to AFS (Andrew File System) that should have better scalability with a cache on the client side and a read-only replication on the server side. The deployment of AFS has been almost done, but the integration with LSF is still in progress.

We set up Quattor (developed by CERN) servers to install OS on most nodes and to manage rpm packages and configuration files on the local facility node. At the Tier-2 site, apt/yum and YAIM are used for package upgrade and gLite configuration, respectively. We prepared Quattor templates by ourselves. The OS installation by Quattor is performed through the PXE (Preboot eXecution Environment).

3.3 Monitor programs

For monitoring purposes, we are using several programs from fabric to Grid services. Some are provided by the vendors and some are made by us. We describe these softwares below.

Lemon (LHC Era Monitoring) is a server/client monitoring system. A sensor is running on each client node and a server collects monitoring information from the clients and visualizes the data through web interface. Because we do not use Oracle as the backend database, an alarm function does not work. To enable alarm, we are going to deploy Nagios which is widely used even in the Grid community.

MRTG (Multi Router Traffic Grapher) and Smokeping are tools for monitoring

network statistics and latency/connectivity, respectively. Both programs use RRDtool (Round Robin Database Tool) internally to store and graph data. MRTG gets statistics data of each port of the Ethernet switches with SNMP (Simple Network Management Protocol). Smokeping is to measure a round trip time (RTT) to designated node or router.

We installed Dell OpenManage on all Dell servers. It checks hardware status and logs events. With IT Assistant from Dell, one can see all events from the web browser. Infortrend provides the RAIDWatch Manager to manage their RAID systems. We can get emails from the RAIDWatch notifying a failure of a hard disk drive or a RAID controller.

The gLite middleware does not provide monitoring tools enough to check the Grid services. Some monitoring information is provided by GStat (using the BDII data) and GridICE (Lemon-like sensor), and the APROC (Asia-Pacific Regional Operations Centre), which supports TOKYO-LCG2 operations, monitors Grid services and network status of our site, but they are still not sufficient for us. We have created some tools to monitor the DPM. The first one is to calculate data transfer rates of the GridFTP servers. It parses GridFTP logs on the disk servers and makes graphs of the data transfer rates per domain name, such as cern.ch. Although it is not very accurate (each log message has only start and end times of a transfer and thus the only average rate can be calculated, and an aborted transfer is not logged), this is very useful, together with the MRTG information of the disk servers. Another home-made tool makes a disk usage table of the DPM. A Perl script runs once a day to retrieve information directly from the MySQL backend database. One can check the disk usage in a directory or owned by a user/group with a web browser.

4. *Network and Data Transfer*

4.1 Network

Network connectivity with other Grid sites is crucial issue, especially at the ICEPP Tier-2 site. In the ATLAS computing model a Tier-2 site must be associated with one Tier-1 site, and the Tier-2's activities (data transfer and job submission) should be mostly limited to that with the associated Tier-1. In our case, it is CC-IN2P3 (Lyon, France) which is very far from Japan and the RTT between the two sites is 280ms.

ASGC (Taipei, Taiwan) is an additional Tier-1 candidate for ICEPP because it is the nearest ATLAS Tier-1 site (RTT is 30ms) and they help Grid operations of our site as the Regional Operation Centre (ROC), as mentioned above. However, the

network link between Tokyo and Taipei (maintained by ASGC) had only 622Mbps bandwidth until last year. It is currently 1Gbps and will be upgraded to 2.4Gbps in the near future. We performed some network and data transfer tests between ASGC and ICEPP, but we have not had formal ATLAS activity with ASGC yet.

In Japan, SINET (Science Information Network) is the major National Research and Education Network (NREN). The bandwidth is 10 to 40Gbps at the backbone, and most ATLAS collaborating institutes in Japan connect to SINET with 1Gbps or more. The ICEPP regional centre connects to SINET with 10Gbps, through the University router. In order not to reduce the available bandwidth, we do not use any IDS or Firewall, but apply an access control list (ACL) at the switches.

SINET provides an international link to New York, where GEANT (European academic network) is also connected. GEANT connects to RENATER (French NREN) to which the CC-IN2P3 connects. Although SINET, GEANT and RENATER are all public lines shared with other traffic, the bandwidth is 10Gbps all the way from ICEPP to CC-IN2P3 since February 2008. Before then, it was limited to 2.4Gbps at the connection between SINET and GEANT routers at New York.

4.2 Data Transfer

It is well known that it is difficult to achieve high performance with TCP over long distance wide-bandwidth network (so-called Long Fat pipe Network, LFN). In TCP, congestion control is realized by changing a window size. Since data transfer rate is roughly given by window size / RTT, the congestion control is critical in order to have high transfer rate. Network condition is also important, because once a packet is lost then TCP window size shrinks and long time is required to recover. This recovering behaviour depends on the TCP implementation. In Linux kernel 2.6 (SLC4) the default TCP implementation is BIC TCP which behaves better than TCP Reno in kernel 2.4 (SLC3) for the congestion control. This is another reason why SLC4 is used on the DPM disk servers.

Another way to improve data transfer rate is to increase number of data streams. In gLite or other Grid middlewares, GridFTP is a standard program used for widearea file transfer. It makes TCP connections in parallel to send a file in multiple streams. Moreover, we can run multiple GridFTP programs concurrently to send multiple files. In the WLCG, data transfer using GridFTP is controlled by a software called File Transfer Service (FTS) in gLite. With FTS, one can set a channel between two SEs, like IN2P3-TOKYO, and set numbers of data streams and concurrent files for a channel. In ATLAS, FTS is usually used from the Distributed Data Management (DDM) system called Don Quijote2 (DQ2) [2].

We set the TCP parameters (window sizes and so on) of the Linux kernel on the

disk servers to modest values (up to 4MB) by using YAIM, but increased number of data streams and number of concurrent file transfers to 10 and 20, respectively, in FTS. These settings have not been fine-tuned but give reasonably good results for data transfer between CC-IN2P3 and ICEPP, without memory exhaustion on the disk servers thanks to many disk servers at both sites.

In order to check network condition, TCP performance was measured with iperf-program between test machines set up at both sites. All the test machines have 1GbE network interface card, hence it is impossible to test 10Gbps WAN bandwidth but still useful to detect unusual condition. TCP window size was able to grow up to 8MB. As a first step, we confirmed that 32bit SLC3 gave worse performance than 32bit SLC4, thus we decided to use 32bit SLC4 boxes afterward. Fig. 2 shows measurement results of iperf in both directions for about 9 days. Numbers of parallel streams was set to 1, 2, 4 and 8 in iperf. From CC-IN2P3 to ICEPP, transfer rates nearly reached the 1Gbps limit of the NIC with 8 streams most of the time, while in the opposite direction it was much worse and unstable. We have not yet understood the reason of the difference, but we are not worried about this asymmetry very much because data transferred from Tier-1 to Tier-2 would be much larger than that from Tier-2 to Tier-1; the former deals with real detector data, but the latter doesn't. However, we hope that the recent network upgrade at New York (2.4Gbps to 10Gbps) has improved performance.

Data transfers in ATLAS has been going on for a long time by using Monte Carlo simulation data or cosmic ray data. In most cases, the ICEPP Tier-2 site receives data from the CC-IN2P3 Tier-1 site. Fig. 3 shows a snapshot of data traffic to the disk servers at ICEPP during a cosmic run in March 2008. It was created by our own monitor program mentioned in 3.3. There were 6 disk servers used at ICEPP and more than 30 disk servers (Solaris, ZFS filesystem) at CC-IN2P3. During the transfer, a peak rate of 140MB/s was observed when both FTS and DQ2 were in good shape.

When we will add 200TB of disk storage (and disk servers) to the Tier-2 site soon, we will also upgrade gLite middleware from 3.0 (32bit) to 3.1 (64bit) on the DPM nodes. GridFTP in gLite 3.1 has been upgraded to version 2 which enhances performance very much. After these upgrades, performance of the data transfer should be improved further.

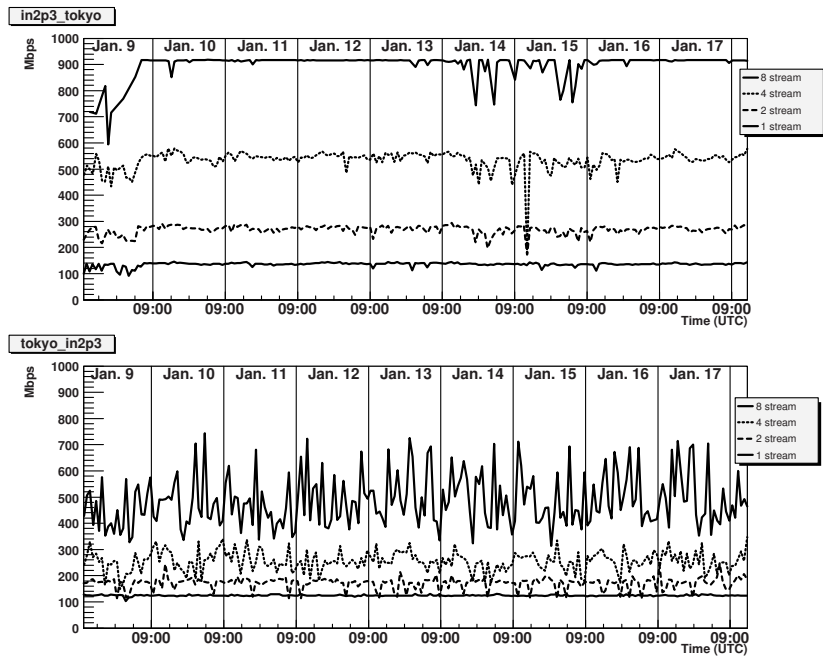


Fig. 2 Network throughputs were measured with iperf program from January 9 to January 17, 2008. Top figure is for CC-IN2P3 to ICEPP, and bottom is for ICEPP to CC-IN2P3. 1Gbps was the absolute maximum of the bandwidth limited at the 1GbE NIC at the both test nodes. We observed bad performance in ICEPP-to-CC-IN2P3 case, but the reason is not understood.

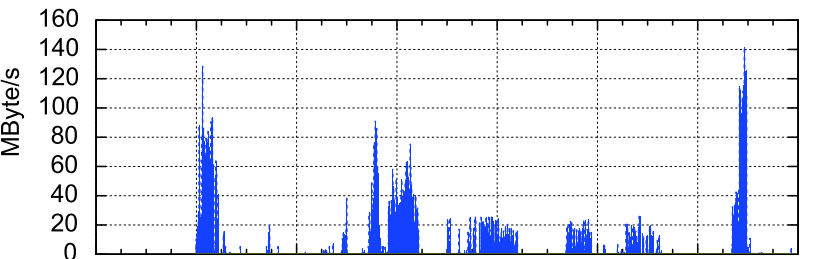


Fig. 3 Data transfer rate from CC-IN2P3 to ICEPP during a cosmic run in March 2008. This graph is created from GridFTP logs at the disk servers. A peak rate of 140MB/s was observed.

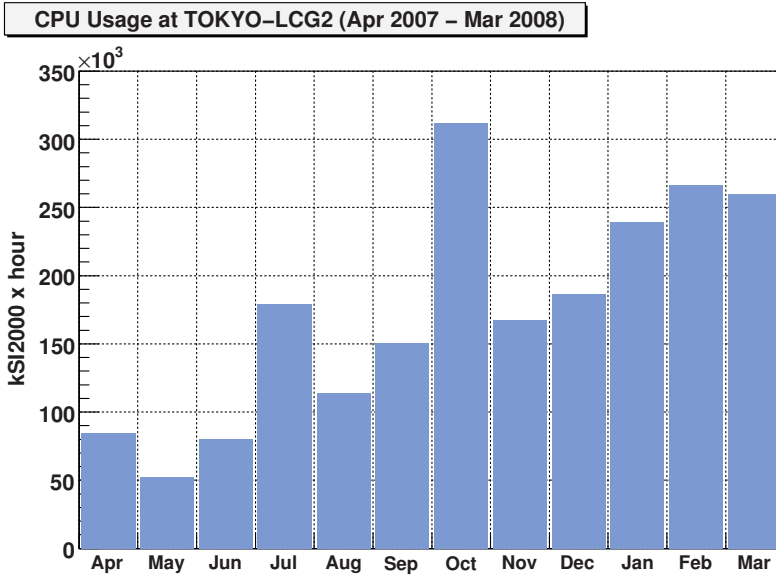


Fig. 4 Monthly CPU usage (in kSI2000_hour) at the ICEPP Tier-2 site from April 2007 to March 2008. Since most jobs are ATLAS Monte Carlo simulations, the usage variation over time depends on the ATLAS activity, although increasing tendency is observed.

5. CPU Usage

CPUs at the Tier-2 site has been used mostly by the ATLAS Monte Carlo simulation. In the last 12 months, from April 2007 to March 2008, 2.09×10^6 (kSI2000xhour) of CPU was used at the ICEPP Tier-2 site. This number is one of the largest contributions to the ATLAS experiment among Tier-2 sites. The usage had increasing tendency as shown in Fig. 4, and it could be increased further as there are still free batch slots from time to time and input data distribution for the Monte Carlo simulation jobs from CC-IN2P3 to ICEPP could be faster in future.

6. Summary

A regional centre for the ATLAS data analysis has been set up at ICEPP, the University of Tokyo. It is comprised of the Tier-2 Grid site and the local facility. Operations of the Tier-2 site have been stable, and availability and reliability measured by WLCG was over 90% recently, with help of monitor programs and use of a testbed. Within the ATLAS activity, data files are transferred from the associated Tier-1 site in France, and the Monte Carlo simulation jobs run on the PC

farms constantly. A peak rate of 140MB/s was achieved recently for the data transfer, and we try to improve the rate further.

The regional centre is operated by 4_5 FTEs. Two system engineers from a company work with the ICEPP staff, mainly on the fabric, and the Tier-2 site is operated by ~1 FTE. This situation will be better once deployment of the new software is completed and the system becomes mature.

Acknowledgements

We would like to thank National Institute of Informatics (NII), network management team of the University of Tokyo, and Computing Research Centre of High Energy Accelerator Research Organization (KEK) for setting up and managing the network infrastructure. We are also grateful to ASGC and CC-IN2P3 staffs for their cooperation and support in performing network tests and data transfer operations.

References

- [1] M. Ishino et al., Nucl. Instr. Meth. A 534, 70 (2004).
- [2] M. Branco et al., J. Phys. Conference Series 119, 062017 (2008).

Production Grids in Asia

Applications, Developments and Global Ties

Lin, S.C.; Yen, E. (Eds.)

2010, XII, 213 p., Hardcover

ISBN: 978-1-4419-0045-6