PUBLICATIONS OF THE KOREAN ASTRONOMICAL SOCIETY
26: 159 ~ 168, 2011

ISSN 1225-1534 (Print)
http://dx.doi.org/10.5303/PKAS.2011.26.4.159

KMTNet 시험운영 서버 구축
CONSTRUCTION OF TEST SERVERS FOR KMTNet DATA MANAGEMENT

김동진, 이충욱, 김승리
한국천문연구원
D. -J. KIM, C. -U. LEE, AND S. -L. KIM
Korea Astronomy and Space Science Institute, Daejeon 305-348, Korea
E-mail: keaton03@kasi.re.kr
(Received 17 November, 2011; Accepted 07 December, 2011)

ABSTRACT

We constructed two test server systems for KMTNet data management. One is the photometry database server which is optimized for stable operation, and the other is the photometric data process server which is optimized for fast I/O between devices. The performances of servers and data storage units were tested using various methods. Database upload was also checked using five different methods. From tests, we concluded that the most efficient method to upload photometric data processing results to database is the use of three nodes with job scheduler under the InnoDB engine. In this study we provide the test results for prototype servers for KMTNet data management.

Key words: instrument; data reduction; KMTNet: database

1. 서론

Wolszczan & Frail(1992)에 의해 태양계 밖에 있는 외계행성의 존재가 최초로 확인된 이후, 2011년 11월 현재 약 700여개의 외계행성이 발견되는(Schneider et al., 2011) 등 천체관측기기의 발달에 따라 외계행성 탐색관측이 국제적으로 매우 활발히 진행되고 있다.

한국천문연구원에서는 미시중력렌즈 방법을 이용한 지구형 외계행성 탐색을 목표로, 2009년부터 외계행성 탐색시스템(Korea Microlensing Telescope Network; KMTNet) 개발 사업을 진행하고 있다. 이 사업에서는 직경 1.6 m 광시야 광학망원경과 3.4억 화소의 모자이크 CCD 카메라로 이루어진 광시야 관측시스템을 제작하고 있으며, 이 시스템을 이용하여 얻어질 대용량의 관측 자료를 전송하고 처리하기 위한 측광 파이프라인 시스템과 측광 데이터베이스 시스템도 함께 개발 중이다.

외계행성 탐색시스템은 2013년과 2014년에 남반구의 칠레, 남아공화국과 호주에 각각 1, 2, 3호기가 건설될 예정이다. 이 시스템으로 얻어질 대용량의 관측 자료는 국내로 전송되어 자료처리 및 분석 과정을 거칠 것이다. 이를 위해 대용량 자료처리 파이프라인과 측광 데이터베이스에 필요한 핵심루틴을 개발했으며(김동진 등, 2009, 2011; Lee et al., 2010), 모의자료를 이용한 시험운영을 통하여 이들 프로그램이 정상 작동하는지 확인할 필요가 있다. 즉, 실제 운영에 사용할 시스템을 사전에 구축하여 우리가 개발한 소프트웨어가 정상 작동하는지, 하드웨어가 제 성능을 내는지, 자료 흐름상의 모든 작업이 정상적인지 확인하여야 한다.

시험운영을 위해서는 소프트웨어 개발단계부터 시스템을 구축해야 하는 비용적인 부담이 있지만, 도입하고자 하는 하드웨어 상호간의 호환성 문제를 파악하고 수치상의 성능이 아닌 실제 체감 성능을 파악하여 소프트웨어 개발에도 참조할 수 있는 장점이 있다. 또한 운영시 발생할 수 있는 다양한 문제점을 미리 확인하여 실제 자료처리가 시작되었을 때 동일한 문제가 발생되지 않도록 할 수 있다.

이 연구에서는 측광 데이터베이스 시스템과 측광 파이프라인 시스템의 시험운영을 위한 시스템을 구축하고, MOA(Microlensing Observations in Astrophysics; Bond et al., 2001) 그룹의 실제 측광자료를 기반으로 생성한 5년간의 가상 측광 파일로 시험운영 서버의 성능이 충분한지 알아보려고 한다. 이를 위해 시험운영 서버의 특성과 성능을 조사하고, 운영체제와의 호환성 문제 및 기타 운영상에 문제가 없는지 평가하였다. 2장에서는 구축한 시험운영 서버, 외부 저장장치, 테이프 백

업장치 및 파일전송 프로그램에 대하여 기술하고, 3장에서는 구축한 시험운영 서버에서 5년간의 측광자료를 저장하였을 경우 데이터베이스의 성능을 기술한다.

2. 시험운영 서버 구성

KMTNet은 3대의 18K × 18K CCD 카메라로 우리 은하 중심방향 4° × 4° 영역을 24시간 연속 관측하며, 하루 동안 생성될 관측 자료의 양은 약 570 GB로 5년간 570 TB의 자료가 생성될 예정이다. 관측 자료는 네트워크나 저장매체를 통하여 한국천문연구원으로 바로 전송 할 것으로, 각 관측소는 16 TB의 저장장치를 설치하여 2달 동안의 자료를 보관하여 자료 전송 시 저장매체의 분실이 발생하면 즉각 대처할 수 있도록 하였다. 이렇게 한국천문연구원에 수집된 관측 자료는 백업과 동시에 측광 자료처리가 진행 될 것이다. 처리된 측광 자료는 데이터베이스에 저장하여 언제든지 검색하여 자료를 분석할 수 있도록 한다. 5년 동안 데이터베이스에 저장할 측광 결과는 약 300 TB 정도가 될 것으로 예측된다.

3대의 관측시스템을 이용하여 24시간 연속 관측을 수행하면, OGLE(Optical Gravitational Lensing Experiment; Udalski et al., 2008)이나 MOA 그룹과 같이, 흥미로운 천체의 후속(follow-up) 관측을 위한 실시간 알람을 제공할 수 있다. 그러나 이를 위해서는 대용량의 자료를 빠른 시간 내에 처리하기 위한 고성능의 시스템이 필수적이다. 이러한 현업 시스템을 운영하기 전에 시스템을 미리 구축하여 성능과 한계를 파악할 필요가 있다. 이렇게 도입한 시험운영 서버는 시험운영이 종료되면 도출된 각종 문제점을 보완하여 현업 장비로 전환한다. 이 장에서는 KMTNet에서 사용할 시험운영 서버의 사양과 기능에 대해서 기술한다.

2.1. 파이프라인 서버와 데이터베이스 서버

KMTNet 측광 파이프라인 시스템과 측광 데이터베이스 시스템은 관측 자료를 빠른 시간에 전송하고 측광 처리하여 데이터베이스에 입력한 후 원하는 결과를 추출할 수 있어야 한다. 두 시스템은 필요할 경우 자원을 공유할 수 있도록 동일한 CPU 성능과 네트워크 전송대역폭을 가지고 있지만 데이터베이스 시스템은 안정성을 기반으로 자료의 업로드 및 검색 속도 등이 KMTNet 프레임워크 개발 서버(김동진 등, 2011)를 기준으로 더욱 빠른 성능을 가지도록 구성하였고, 파이프라인 시스템은 네트워크를 통한 관측 이미지 전송 및 저장이 기존의 측광파이프라인 개발 서버(이충욱 등, 2009)보다 빠른 성능을 가지도록 구성하였다. 그림 1은 이번 연구를 위해 구성한 KMTNet 시험운영 서버의 모습이다.

측광 데이터베이스 시스템은 데이터베이스가 작동하는 동안 순간정전과 같은 돌발 상황을 방지하기 위해

그림 1. KMTNet 시험운영 서버. 3대의 데이터베이스 서버와 3대의 파이프라인 서버, 16 TB 저장장치와 LTO-5 테이프로 구성되었다.

전원을 2중화 하였다. 시스템이 갑작스레 꺼질 경우 데이터베이스 파일이 손상될 수 있고 복구에 상당한 시간을 소비해야 하기 때문에 안정적인 전원 공급이 필요하다. 따라서 데이터베이스 시스템에 전원공급장치(Power Supply Unit; PSU)를 2개씩 장착하고, 각각의 PSU에는 독립된 차단기에서 분배된 전원이 인가되도록 하였다. PSU에 무정전 전원장치(Uninterruptible Power Supply; UPS)를 연결하면 순간 정전이나 장시간 정전이 발생하였을 경우 서버 다운으로 인한 데이터베이스 손상을 막을 수 있다. UPS는 관측시스템 1호기가 완성되고 시험관측이 들어가기 전에 장착할 계획이다. 그리고 정전이 발생하거나 정전 이외 문제로 서버의 정상 운영이 불가능할 경우 외부에서 원격으로 접속하여 전원을 끄고 내리거나 문제점을 수정할 수 있는 독립된 네트워크 포트를 확보하여 시스템에 포함하였다. 이는 운영체제의 상태와 서버의 전원 인가 여부에 관계없이 접속하여 복구작업을 할 수 있도록 돕기 위함이다.

측광 파이프라인 시스템은 안정성 보다는 자료의 입출력 속도 향상에 중점을 두고 구성하였다. 즉 640 MB 크기의 관측 이미지 1장을 처리하는 과정에 다른 서버로 파일을 전송할 때 사용되는 HDD와 네트워크의 입

표 1. 측광 데이터베이스 서버와 측광 파이프라인 서버의 사양

	측광 데이터베이스 서버	측광 파이프라인 서버
CPU	Intel 6-Core Xeon X5650 2.66 GHz × 2	Intel 6-Core Xeon X5650 2.66 GHz × 2
노드 수	3	3
RAM	12 GB/node	24 GB/node
HDD	140 GB (SAS 6 Gbps, 10 KRPM)	300 GB (SAS 6 Gbps, 10 KRPM)
POWER SUPPLY	750 W × 2 (Master node) 450 W × 2 (2 Slave node)	450 W
OS	Scientific Linux[1] 6.1	Scientific Linux 6.1
성능 목표	시스템 안정성, 자료입력 및 검색 속도 향상	자료 전송 및 저장 속도 향상

표 2. HDD와 SSD의 쓰기 속도 비교(828 MB)

저장장치	CentOS[2] 5.7 Linux, 64 bit	Windows 7, 64 bit
HDD → HDD	1.1 s	12 s
HDD → SSD	8.9 s	19 s

출력 성능을 높이는데 중점을 뒀다. HDD는 Serial Attached SCSI(SAS) 인터페이스와 6 Gbps 전송대역폭을 사용하고 10 K RPM으로 고속 회전하는 제품을 장착하여 자료의 접근속도 및 전송속도를 증가시켰다. SAS 인터페이스는 여러 개의 HDD와 인터페이스 카드를 병렬로 연결하던 기존의 Small Computer Small Interface(SCSI)를 직렬방식으로 연결하여 병목현상을 개선한 인터페이스로 개인용 PC보다는 서버에 주로 사용된다. 표 1에 측광 데이터베이스 서버와 측광 파이프라인 서버의 사양을 정리하였다.

특히 서버 구성에 일반 HDD를 사용하였는데, SSD를 사용하지 않고 10 KRPM의 고회전 HDD를 선택한 것은 시험 결과 SSD의 경우 자료를 읽는 속도는 빠르지만 HDD보다 쓰는 속도가 느려 쓰기 작업이 많은 자료처리에서는 SSD가 큰 이점이 없기 때문이다. 표 2는 리눅스와 윈도우 시스템에서 828 MB의 자료를 SSD와

[1] Scientific Linux는 페르미 연구소와 유럽핵입자물리연구소 및 기타 대학과 연구기관에 의해 개발, 배포되는 무료 배포판으로 RehHat Enterprise Linux와 호환된다.

[2] CentOS Linux는 자체 커뮤니티에 의해 관리되는 운영체제로 RehHat Enterprise Linux의 소스 코드를 이용하여 개발 및 배포된다.

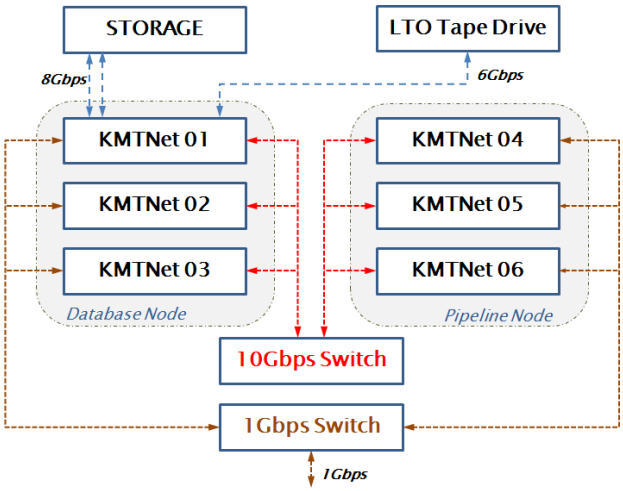

그림 2. 시험운영 서버와 네트워크 구성도. 저장장치는 8 Gbps Fiber Channel, LTO 테이프 드라이브는 6 Gbps로 연결하였고, 서버는 10 Gbps 네트워크로 연결하였다.

HDD(7.2 KRPM)에 저장할 때 소요된 시간으로, 실험에 사용한 모든 저장장치는 3 Gbps의 포트에 연결하였다. 이 실험으로 운영체제에 따라 속도의 차이는 있지만 SSD의 쓰기 속도가 HDD보다 느린 것을 볼 수 있다.

시험운영 서버에 사용할 네트워크로 내부망은 10 Gbps, 외부망은 1 Gbps로 연결하였다. 내부망은 주로 파일 전송과 외부 저장장치를 Network File System(NFS)로 연결하는데 사용된다. 그림 2는 시험운영 서버와 네트워크 구성도이다.

2.2. *외부 저장장치*

KMTNet의 핵심 연구주제인 중력렌즈 현상을 이용한 외계행성 탐색에 사용되는 총 저장용량은 수 PetaByte(PB)에 이른다. 이러한 저장용량은 그 양이 너무 방대하여 데이터베이스 시스템이나 파이프라인 시스템 자체에 HDD를 부착하여 구성하기는 매우 어렵기 때문에 별도의 저장장치를 구성한 후 SAS, Fiber Channel(FC) 또는 iSCSI 인터페이스 등으로 연결한다. 한편 외부 저장장치를 구축할 때는 저장용량의 확장성, 자료 흐름의 병목현상 및 여러 서버가 하나의 저장장치에 연결되었을 경우 발생하는 파일의 신뢰성 등이 충분히 검토되고 고려되어야 한다.

특히 이번 연구에서는 외부 저장장치 구성을 위해 HDD를 한꺼번에 구입하지 않고, 필요할 때마다 용량을 증가시킬 수 있는 방식을 도입하여 확장성을 높였다. 또한 확장성 이외에도 외부 저장장치에는 수십 ~ 수백 개의 HDD를 장착하여 RAID로 구성하기 때문에 컨트

그림 3. 외부 저장장치 및 인터페이스 모듈. 총 12개의 HDD를 장착할 수 있고, 2개의 컨트롤러 모듈을 장착하여 2중화 구성을 하였다.

롤러의 성능과 서버와 연결하기 위한 인터페이스 지원도 충분히 고려되어야 한다.

이번 구성은 시험운영을 위한 시스템 구축이므로 최소 단위 저장장치 개념을 도입하였다. 최소 단위 저장장치는 인클로저 단위로 확장이 가능하고, 하나의 인클로저에 6 Gbps 전송대역폭을 지원하는 12개의 7.2 KRPM 2 TB HDD를 장착하여 RAID 5로 구성하였다. 인클로저에는 2개의 컨트롤러 및 인터페이스 모듈을 장착하여 확장 시 발생할 수 있는 문제 해결능력을 높였다. 2개의 컨트롤러는 failover로 설정되어 A 컨트롤러가 작동 중 하드웨어 고장을 일으키면 B 컨트롤러는 대기 상태에서 깨어나 A 컨트롤러 기능을 그대로 넘겨받게 된다. 컨트롤러뿐만 아니라 FC로 연결된 인터페이스도 이중화로 구성하여 failover로 작동한다. 그리고 독립된 전원으로 2중 구성하여 하드웨어 문제나 정전에 의한 자료 손실을 최소화하였다. 인클로저는 최대 8대의 단위 저장 장치까지 확장 설치할 수 있어 최대 96개의 HDD를 장착할 수 있다.

서버와 외부 저장장치는 8 Gbps를 지원하는 FC 인터페이스로 구성하였고, 서버에는 PCI-Ex(ver 2.0) 8 x 전송대역폭(양방향 8 Gbps)을 사용하는 Host Bus Adapter(HBA)를 장착하여 자료 전송시 병목현상이 발생하지 않도록 하였다. 그림 3은 도입한 단위 저장장치로 총 16 TB의 저장용량을 갖는다.

2.3. 개방선형테이프(LTO)와 ASPERA P2P

남반구 관측소에서 얻은 관측 자료는 천문연구원으로 옮겨져서 처리·분석될 계획이다. KMTNet 망원경은 칠레, 남아프리카 공화국, 호주에 설치되며, 칠레를 제외한 2곳은 국제 백본망 또는 현지 관측소의 네트워크 사정이 열악하기 때문에 관측 자료를 항공우편을 통해 수집할 계획이다. 또한 칠레는 네트워크 전송이 불가능할 경우를 대비한 비상 계획도 함께 수립하여야 한다. 따라서 관측 자료를 어떤 저장매체에 담아서 보낼지에 대한 충분한 검토가 이루어져야 한다. 현재 널리 사용되는 저장매체로 HDD와 테이프의 두 가지를 고려할 수 있는데 HDD는 전송 속도가 빠르고 PC에 쉽게 연결할 수 있다는 장점이 있지만, 고속의 구동부와 전자부품이 포함되어 있어 충격에 취약하고 무겁다는 단점이 있다. 테이프는 크기가 작아 이동이 쉽고 HDD 무게의 1/3로 가벼운 장점이 있지만, 자료를 읽기 위한 전용 테이프 드라이브가 필요하고 저장 속도가 느린 단점이 있다. 우리는 KMTNet에서 사용할 저장매체로 이 두 가지의 장단점을 보완한 LTO를 사용하기로 하였다.

개방선형테이프(Linear Tape-Open; LTO)는 고속 데이터 처리와 대용량 형식으로 만들어진 백업용 테이프로 자동 오류보정 장치, 오류정정 코드, 하드웨어 데이터 압축, 트랙 레이아웃 기능을 가지고 있다. 2개의 릴이 달려있는 기존의 DAT 테이프 카트리지 형식과 다르게 1개의 릴만 있으며, 최근에는 LTO-4와 LTO-5 규격이 사용된다.

KMTNet 시험운영 서버에서 관측자료 백업 및 자료 전송용으로 도입한 LTO-5 테이프 드라이브는 1.5 TB (압축 저장 시 3 TB)의 저장 공간과 140 MB/s의 자료 전송 속도를 제공하며 LTFS(Linear Tape File System)의 파일 시스템을 지원한다. 6 Gbps의 SAS 인터페이스로 연결하여 테이프 드라이브와 서버간의 자료전송대역을 충분히 확보하였다.

LTFS는 LTO 드라이브 및 미디어 테이프를 일반적인 디스크 파일 시스템과 유사하게 사용할 수 있도록 테이프 저장장치를 파티셔닝하는 것으로, tar 명령을 이용하여 순차적으로 자료를 기록하는 전통적인 테이프 백업 방식과 다르게 HDD와 같이 독립된 저장 공간으로 인식시켜 ls, cp, mv, rm 등의 리눅스 명령으로 자료를 관리할 수 있다. LTFS로 인식된 LTO 테이프는 크게 두 개의 파티션 구조로 이루어진다. 첫 번째 인덱스 파티션에는 파일에 대한 인덱스 정보가 들어가고, 두 번째 데이터 파티션에는 파일이 들어가는 구조이다. 파일에 대한 읽기 요청이 들어오면 첫 번째 파티션의 인덱스 정보를 통해 파일의 위치를 찾아내고, 두 번째 데이터 파티션에서 실제 파일을 찾아 읽어오게 된다. 파일을 삭제할 경우 데이터 파티션의 자료를 삭제하지 않고 인덱스 파티션에서의 정보만 삭제하게 된다. 그리고 파일을 덮어쓰는 경우 HDD와 달리 기존 파일이 있던 자리 위에 덮어쓰지 않고 자료가 저장된 마지막 지점에 저장하게 되므로 일반 HDD처럼 읽고 쓰는 작업이 많을 때는 LTO 테이프가 적합하지 않지만, KMTNet처럼 단순한 파일의 복사에는 Write Once Read Many(WORM) 형태의 LTO 테이프가 적합하다.

관측소의 네트워크 백본망이 잘 구축되어 있다면 네

표 3. 개발 시스템에 따른 I/O 벤치마크(sysbench, ext4)

구분	KMTNet 프레임워크 개발시스템	데이터베이스 (시험운영서버)	측광파이프라인 (시험운영서버)
초당 요청개수 (Requests/s)	78.93	273	393
초당 처리량 (MB/s)	1.23	4.26	6.14

트워크로 자료를 전송하는 방법을 사용할 수 있다. 네트워크로 전송하면 받는 즉시 관측 자료를 처리하여 결과를 볼 수 있는 장점이 있다.

네트워크로 자료를 가져오기 위해서는 일반적으로 FTP 프로토콜을 사용한다. FTP는 21번 포트를 사용하며 자료전송 시 패킷 단위로 데이터를 묶어서 전송하며 보낸 패킷이 정상적으로 수신되면 다음 패킷을 보내게 된다. 만약 수신 과정에서 에러가 검출되면 해당 패킷을 재송신하라고 명령하게 된다. 이러한 과정을 통해 에러가 없는 정상적인 파일을 수신할 수 있다. 그러나 이런 과정을 거치면서 시간적인 손실이 발생하고 전송대역을 전부 사용하지 못하는 문제점이 발생한다. UDP 프로토콜의 경우 FTP와 다르게 에러 보정기능이 없기 때문에 파일 전송을 매우 빠르게 할 수 있다는 장점이 있지만, 전송하는 도중에 에러가 발생하면 손상된 부분을 재전송하지 않아 전송이 완료된 파일은 손상된 상태로 있게 된다.

KMTNet에서 도입한 ASPERA P2P는 UDP의 단점을 보완하기 위해 UDP 프로토콜에 에러보정 기능을 추가한 프로그램으로 백본망이 좋다면 지정한 전송대역폭을 모두 활용하여 자료를 보낼 수 있다. ASPERA P2P는 서버와 클라이언트가 1 : 1 접속으로 이루어지며 최대 45 Mbps의 전송대역폭을 사용할 수 있다. 한 관측소에서 하루 동안 약 150 GB의 관측 자료에 대해 45 Mbps 전송대역폭을 모두 사용하면 8시간 만에 자료를 모두 가져올 수 있다.

ASPERA P2P를 사용할 경우 지진으로 인한 국제 백본망 손상이나 현지 관측소의 문제로 파일 전송이 불가능한 경우를 대비한 비상계획을 수립하여야 한다. 각 관측소에는 16 TB 용량의 저장장치가 설치되어 2달 이상 관측 자료를 보관할 수 있다. 장비 문제 등으로 인해 네트워크 회선의 단기간 속도저하 또는 접속불량이 발생할 경우 해당 자료는 네트워크가 정상화 된 후 재전송을 할 수 있다. 만약 지진 등으로 국제 백본망이 끊어져 대륙 간 우회를 해야 하는 상황이 발생하면 장기간 네트워크를 통해 자료를 받는 것이 매우 어려워진다. 그래서 칠레 관측소에는 측광 자료처리 서버를 설치하여 산출된 결과 파일을 네트워크로 보내고 관측 자료는 LTO-5 테이프로 보내는 방법을 고려하고 있다.

표 4. 시험운영서버 저장장치에 따른 I/O 벤치마크(sysbench)

구분	내부 저장장치 (Database Node)	외부 저장장치 (Fiber Channel)	외부 저장장치 (10 Gbps/NFS)
초당 요청개수 (Requests/s)	273	253	239
초당 처리량 (MB/s)	4.26	3.95	3.74

3. 시험운영 서버 성능 실험

본 연구를 위해 구성한 시험운영 서버 및 저장장치의 성능을 확인하기 위하여 자료 입출력, 자료백업 및 전송, 데이터베이스 운영 등의 실험을 실시하였고, 구성한 시스템의 성능을 평가하였다.

3.1. 자료 입출력 실험

시험운영 서버의 입출력 속도를 측정하기 위하여 sysbench 프로그램을 이용하였다. 이 프로그램은 MySQL 데이터베이스 라이브러리와 연동하여 특정한 크기의 임시데이터를 생성한 후 무작위로 읽기/쓰기를 반복하여 저장장치가 초당 처리할 수 있는 작업의 요청수와 초당 처리량이 얼마나 되는지 확인할 수 있다. 표 3은 KMTNet 프레임워크 개발 시 사용한 시스템(김동진 등, 2011)과 시험운영 서버의 벤치마크 결과로, ext4 파일 시스템을 기반으로 테스트하였다.

입출력 속도 시험 결과 측광 데이터베이스 시스템은 프레임워크 개발 서버보다 약 4배 이상 빠르고, 측광 파이프라인 서버는 6배가 빠른 것을 확인하였다. 이와 함께 저장장치의 종류에 따른 벤치마크 결과를 표 4에 정리하였다. 외부 저장장치의 속도는 서버의 HDD에 비해 약 10% 정도 낮은 속도를 보여준다. 이는 외부 저장장치의 경우 전송대역폭이 서버의 HDD와 비슷해도 7.2 KRPM의 회전속도를 가지는 HDD를 사용했기 때문에 자료를 읽고 쓰는 속도가 느린 까닭으로 여겨진다. 한편 외부 저장장치의 경우 6대의 시험운영 서버 중 1대만 FC 인터페이스로 연결하였고 다른 서버는 10 Gbps 네트워크 기반에 NFS로 연결하여 충분한 대역폭을 확보하였기 때문에 NFS로도 FC 인터페이스와 비교하여 비슷한 속도를 확보하였다.

네트워크 속도에 따른 자료의 전송속도를 표 5에 정리하였다. scp 명령어로 828 MB의 파일을 1 Gbps와 10 Gbps 네트워크에서 각각 전송하였을 경우 특별한 속도 차이는 발생하지 않았다. scp는 암호화 기능이 포함된 복사 명령어로 다른 서버로 자료전송 시 암호화 하는 과정으로 인해 고속 네트워크에서 전송속도의 향상이

표 5. 네트워크 따른 전송속도 벤치마크(828 MB)

네트워크 속도	scp 명령어	ftp/ncftp 명령어
1 Gbps	18.00 s	7.43 s
10 Gbps	18.00 s	1.41 s

없었다. 그러나 ftp나 ncftp와 같이 일반 프로토콜을 사용할 경우 10 Gbps의 네트워크가 1 Gbps보다 전송속도에 있어 6배 이상 빨랐다. 동일한 파일을 10 KRPM HDD에서 복사할 경우 0.7 ~ 0.8초가 소요되는 것에 비교하면 10 Gbps의 경우 HDD 속도에 근접하는 것을 볼 수 있다. 즉 10 Gbps 네트워크를 사용한다면 자료를 전송할 때 발생하는 시간 지연을 충분히 해결할 수 있을 것이다. KMTNet 시험운영 서버의 내부 자료전송은 스크립트 구성이 가능한 ncftp와 10 Gbps 네트워크를 사용하는 것이 최선이라 판단된다.

표 6은 기존 측광 파이프라인 개발 시스템과 시험운영 측광 파이프라인 시스템과의 네트워크 파일 전송 및 복사 실험의 결과이다. 시험운영 측광파이프라인 시스템은 기존 시스템과 비교하여 파일 전송은 8배, 복사는 5배 빠른 결과를 보였고, 시스템의 구축 성능 목표를 충분히 만족하였다.

3.2. 개방선형테이프(LTO) 전송속도

KMTNet 관측소에서 관측 자료를 보내기 위해 사용할 LTO-5 테이프와 HDD의 자료 백업 속도를 비교하였다. 한 관측소에서 하룻밤에 생성되는 관측 자료는 약 190 GB이지만 약 10% 더 많은 용량인 203 GB로 가정하여 모의자료를 생성하였고, 이 자료를 외부 저장장치에서 백업 미디어로 복사하는 과정을 각각 5회 반복하였다. 외부 저장장치는 8 Gbps FC 인터페이스로 연결하였고, LTO-5 테이프는 6 Gbps SAS 인터페이스로 연결하였다. LTO-5와 비교할 백업 미디어로는 SATA-2 HDD(3 Gbps, 7.2 KRPM, 1 TB)을 사용하였고 e-SATA 인터페이스로 연결하였다. 표 7은 LTO-5 테이프와 HDD의 저장시간 측정 결과이다.

리눅스의 cp 명령으로 외부 저장장치에서 HDD로 자료를 복사할 경우 평균 39.34분의 시간이 소요되었고, HDD에서 외부 저장장치로 자료를 다시 복사할 경우 33.34분의 시간이 소요되었다. 외부 저장장치에서 LTFS 파일시스템을 사용한 LTO-5 테이프로 자료를 복사할 경우 51.22분의 시간이 소요되었으며, LTO-5 테이프에서 외부 저장장치로 복사할 경우 37.41분이 소요되었다. 기존 백업방식인 tar 명령어로 LTO-5 테이프에 복사하

표 6. 측광파이프라인 시스템에 따른 관측이미지의 파일 전송 및 복사 실험(828 MB)

구분	측광파이프라인 개발시스템 (2009)	측광파이프라인 (시험운영서버)
파일 전송 (ftp/ncftp)	11 s (1 Gbps사용)	1.4 s (10 Gbps 사용)
파일 복사 (cp)	8.71 s	1.7 s

표 7. 외부 저장장치와 LTO-5 테이프·HDD의 자료 저장 속도

저장매체	저장시간	속도
외부 저장장치 → SATA-2 HDD (7,200 RPM)	39.34분	89.12 MB/s
SATA-2 HDD (7,200 RPM) → 외부 저장장치	33.34분	105.95 MB/s
외부 저장장치 → LTO-5 테이프	51.22분	67/65 MB/s
LTO-5 테이프 → 외부 저장장치	37.41분	91.90 MB/s

였을 경우 전송속도가 약 30 MB/s로 100분 이상의 시간이 소요되었다.

자료를 외부 저장장치에서 HDD와 LTO-5로 저장할 경우 LTO-5 테이프가 HDD보다 약 25% 속도가 떨어지는 것을 확인할 수 있다. 하지만 백업미디어에서 다시 자료를 불러오는 경우에는 HDD가 LTO-5 테이프에 비해 약 10% 정도 빠른 결과를 보였다. 이것은 LTO-5 테이프의 자료 저장 속도가 HDD에 비해 현저히 떨어지지 않고, 읽기 성능은 HDD와 거의 동일하다는 것을 의미한다. 그러므로 LTO-5 자체의 장점을 고려하면, HDD를 대체하여 현지 관측소에서 자료를 백업하여 발송하기 위한 미디어로 LTO-5가 매우 적합하다고 판단된다.

3.3. 데이터베이스 실험

우리가 구성하고자 하는 측광 데이터베이스 시스템은 자료 저장과 동시에 검색 서비스까지 수행해야하기 때문에 하나의 테이블을 이용하여 시스템을 구성하는 것은 불가능하다. 따라서 이를 해결하기 위해서는 데이터베이스 시스템을 병렬화 또는 분산화 하는 작업을 거쳐야 한다. 병렬화는 데이터베이스 기능을 복수의 서버가 하나의 시스템처럼 작동하도록 공유하는 방식으로 핫백업 기능이 있기 때문에 한 대의 서버가 다운되더라도 데이터베이스 기능에는 큰 문제가 발생하지 않는다. 그

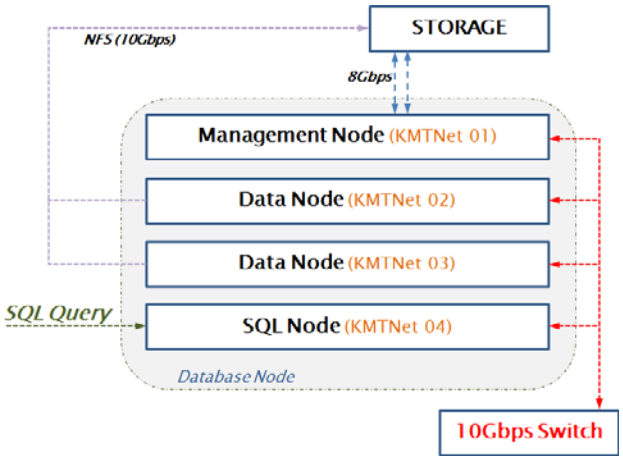

그림 4. MySQL 병렬데이터베이스 구성도. 관리 노드 1대, 데이터 노드 2대, SQL 노드 1대로 구성하였고, 10 Gbps 네트워크로 연결하였다.

러나 분산화는 데이터베이스의 테이블을 여러 서버로 분할하는 방식이기 때문에 한 대의 서버가 다운될 경우 그 서버가 가지고 있는 자료는 검색할 수 없다.

병렬화를 지원하는 데이터베이스는 ORACLE, MySQL, MS-SQL, CUBRID 등 여러 제품이 존재하지만, 김동진 등(2011)이 KMTNet 프레임워크 개발에 사용하였던 MySQL을 이용하여 병렬 데이터베이스 구성을 하였다. 시험운영을 위해 구성한 시스템의 MySQL 병렬 데이터베이스를 구성하기 위해서는 최소 4대의 서버가 필요하다. 모든 노드의 상태 및 기능 전환을 지시하기 위한 관리 노드 1대와 입력되는 자료가 저장될 데이터 노드 2대, 검색 질의 및 자료 입력 명령이 수행되는 SQL 노드 1대 등이다. 규모에 따라 데이터 노드와 SQL 노드를 추가할 수 있고 최대 63대까지 확장할 수 있다. 그림 4는 이 실험에서 구성한 MySQL의 병렬 데이터베이스 시스템 구성도이다.

MySQL 병렬 데이터베이스의 특징은 병렬 구성 전용 검색엔진인 NBD Cluster를 사용해야 하고, 데이터 노드에 저장되는 자료는 스토리지 공유 방식이 아닌 백업 방식으로 작동된다. 자료를 입력할 경우 각각의 데이터 노드에 동일한 자료가 입력되기 때문에 하나의 데이터 노드가 다운되어도 다른 데이터 노드에서 동일한

[3] NDB Cluster는 MySQL의 병렬클러스터를 지원하는 저장엔진으로, 중복성과 부하분산 기능을 하도록 생성되었고, 트랜잭션 기능을 지원한다. 데이터 노드의 데이터를 공유하지 않는다.

[4] InnoDB는 트랜잭션을 처리하기 위해 개발된 저장엔진으로 안전성과 자동장애복구 기능이 우수하다.

[5] MySQL에 포함된 가장 오래된 저장엔진으로 빠른 입출력 및 검색 속도를 가지나, 트랜잭션 기능을 지원하지 않는다.

표 8. 저장엔진에 따른 자료 입력시간(단위: 초)

	레코드 수	NDB Cluster[3]	InnoDB[4]	MyISAM[5]
1 - 30일	62,004,960	25,164.37	12,950.11	348.13
31 - 60일	62,004,960	25,634.05	13,290.72	946.88
61 - 90일	62,004,960	25,705.24	13,268.79	1,073.67
91 - 120일	62,004,960	25,600.21	13,345.05	979.56
121 - 150일	62,004,960	25,837.46	13,536.78	976.48

서비스가 가능하지만 저장 공간이 데이터 노드 수만큼 요구되기 때문에 병렬 데이터베이스로 KMTNet을 장기간 운영할 경우 외부 저장장치의 구입비용은 데이터 노드의 수만큼 증가하게 된다.

우리가 구성한 시스템을 이용하여 데이터베이스 저장엔진에 따른 자료 입력시간을 실험하였다. 실험에 사용한 자료는 MOA 그룹이 2006년 3월 29일부터 2007년 9월 16일까지 관측한 R-band 자료를 사용하였다(김동진 등, 2011). 2K × 4K의 영상을 32개의 512 × 512 영역으로 분할한 이미지 중 하나를 측광처리한 후 RANDOM 함수를 이용하여 5년 동안의 가상 자료를 생성하였다. 이 자료는 1년 중 약 200일 동안 관측을 수행하고, 512 × 512 영역에 14,353개의 별이 있다고 가정하였다. 영상을 512 × 512 크기로 자른 이유는 관측 자료의 왜곡 효과 등에 의한 영향을 받지 않는 최적의 크기라 판단했기 때문이다. 표 8은 모의자료를 30일씩 묶어 150일 동안의 자료를 데이터베이스에 입력한 결과이다. 입력에 사용된 30일치의 자료는 6,200만 레코드를 가지고 있고 크기는 6.2 GB이다.

MyISAM 저장엔진은 업로드 수행시간이 20분 미만으로 가장 빨랐고, InnoDB의 입력 속도는 3시간 40분이었다. NDB Cluster는 SQL 노드에서 데이터 노드로 연결된 네트워크를 통하여 자료가 저장되었고, 2대의 데이터 노드에 각각 저장되기 때문에 7시간이 넘는 시간이 소요되어 병렬화를 수행했을 때 시간이 많이 소요되었다. 병렬화의 장점은 안정성과 로드 밸런싱 능력이다. 그러나 자료 저장에 많은 시간이 소요되어 제때 자료를 데이터베이스에 올리지 못하고 활용하지 못한다면 이러한 장점은 큰 의미가 없다. 그러므로 KMTNet의 측광 데이터베이스는 병렬화보다 관측 영역을 분할하여 데이터베이스를 분산하는 것이 더 효율적이라고 판단된다.

KMTNet에서 사용할 18K × 18K CCD 카메라의 관측 영상을 512 × 512 영역으로 분할하면 총 1,296조각이 된다. 이 조각 중 하나를 측광처리하면 약 1.5 MB의 측광결과 파일이 생성되므로, 18K × 18K 전체 영역에 대한 측광결과 파일 크기는 약 1.9 GB가 된다. 1.9 GB의 측광 자료를 데이터베이스에 입력하는 5가지 경우에 대한 실험을 수행하였고, 표 9에 5가지 실험의 결

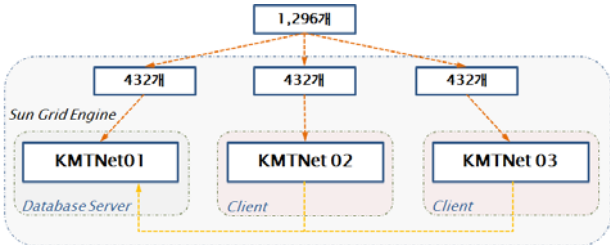

그림 5. 다중 노드에서 job scheduler를 이용한 자료 입력. 1,296개의 명령을 432개씩 3등분하였고, job scheduler를 이용하여 3대의 데이터베이스 서버에서 각각 실행하였다.

과를 정리하였다.

CASE 1과 CASE 2는 1대의 데이터베이스에서 1,296조각의 자료를 입력하는 실험이다. 이중 CASE 1은 데이터베이스 서버에서 job scheduler 프로그램인 Sun Grid Engine으로 1,296회의 입력 명령을 job queue에 올려 자료를 입력한 실험이고, CASE 2는 1,296회의 입력 명령을 단순 반복한 실험이다. 두 경우는 job scheduler의 유무 차이만 존재하는데 MyISAM 저장엔진의 경우 job scheduler를 사용하였을 경우 약 130% 이상의 시간이 더 소요되었고 InnoDB 저장엔진은 큰 차이가 없었다. MyISAM 저장엔진이 한 조각의 자료를 데이터베이스에 입력하는데 1초미만의 시간이 소요되고 InnoDB 저장엔진은 3초가 소요된다. MyISAM과 job scheduler가 함께 사용되면 자료를 입력하는 시간보다 job scheduler가 다음 제어명령을 수행하는 시간이 더 길어 입력속도가 느려진다. 반면 InnoDB는 자료를 입력하는 시간동안 job scheduler가 제어명령을 충분히 처리하므로 입력속도에 큰 차이가 없다.

CASE 3에서 CASE 5까지 실험은 1대의 데이터베이스와 2대의 클라이언트를 이용하여 1,296조각을 분할하여 입력하는 실험이다. CASE 3은 1,296조각을 job queue에 한꺼번에 올린 후 각 서버로 작업을 분배하여 입력한 실험으로, CASE 1과 실험 방법이 동일하다. 그러나 job scheduler가 각 서버의 여유자원을 파악하여 작업을 분배하기 때문에 모든 서버에 동일한 수의 작업이 분배되지는 않는다. CASE 3은 CASE 1보다 MyISAM은 입력 시간이 35% 정도 감소되었고, InnoDB도 약 28% 감소하였다. 그러나 MyISAM은 여전히 job scheduler를 사용하지 않은 CASE 2보다 입력 시간이 오래 걸렸다.

CASE 4는 1,296개의 조각을 432개로 분할하여 ssh 명령어로 각 서버가 독립적으로 입력을 실행하도록 하는 방법으로 CASE 2를 3대로 분산한 경우이다. CASE 2에 비해 MyISAM은 55%, InnoDB는 75% 이상 입력 시간이 감소하였다.

CASE 5는 CASE 4의 실험에 job scheduler를 이용하

표 9. 1,296조각의 측광 자료 입력시간(단위: 초)

저장엔진	CASE 1 **1 node (job scheduler)**	CASE 2 **1 node**	CASE 3 **3 node (job scheduler)**	CASE 4 **3 node (ssh)**	CASE 5 **3 node (ssh + job scheduler)**
MyISAM	350.83	144.87	231.96	65.89	45.75
InnoDB	640.25	645.34	465.88	160.36	115.28

였다. CASE 3의 실험은 1,296조각 모두 job queue에 올리는데 비해 CASE 5는 432조각을 실행하도록 하는 명령어 3개를 queue에 올리므로 job scheduler의 제어시간을 크게 사용하지 않았다.

CASE 5는 CASE 2에 비해 MyISAM은 68%, InnoDB는 82% 이상 입력 시간이 감소하였다. 그림 5는 CASE 5 방법으로 자료를 입력하는 흐름도이다.

MyISAM 저장엔진은 InnoDB 저장엔진과 다르게 하드웨어에 크게 의존하지 않아 빠른 속도로 사용이 가능하다(김동진 등, 2011). 그러나 자료를 입력하는 동안 테이블 전체에 잠금을 설정하고 작업하기 때문에 검색이 불가능하고, 반대로 검색을 하는 동안에는 자료 입력이 불가능하다. 데이터베이스에 저장된 측광자료에서 변광천체를 검출하기 위한 프로그램이 데이터베이스에 연동되어 24시간 실행되는 것을 고려한다면 InnoDB 검색엔진을 사용하는 것이 바람직하다. 그러나 InnoDB의 여러 장점에도 불구하고 자료 입력 시간이 오래 걸리는 단점 때문에 실제 적용에 어려움이 있었지만, CASE 5의 방법으로 자료를 입력한다면 InnoDB 검색엔진을 사용할 때 발생되는 입력시간 지연 문제와 검색 프로그램의 동시 실행을 모두 해결할 수 있으리라 판단된다.

데이터베이스 실험을 위해 생성한 512 × 512 영역의 5년간의 가상자료를 데이터베이스에 모두 입력하면 약 21억 개의 레코드가 생성되고, 용량은 140 GB정도이다. 대용량의 자료를 빠르게 검색하기 위해서 인덱스를 사용하였다(김동진 등, 2011). 표 10은 검색엔진과 인덱스 유무에 따른 검색 소요시간이다. 인덱스를 사용한 MyISAM 저장엔진과 InnoDB 저장엔진에 따른 검색시간의 차이는 없었다. 데이터베이스 테이블에 최초 검색 명령을 실행하면 전체 테이블 정보를 메모리로 캐싱하는 작업이 이루어져 검색 시간이 증가하였지만 그 이후 검색부터 1초미만의 시간이 소요되었다.

KMTNet 프레임워크 개발 시스템의 경우 2,500만개의 자료를 사용하여 실험을 진행하였지만, 이번 연구에서는 그보다 100배 많은 자료를 이용하여 실험하였다. 김동진 등(2011)에 따르면 KMTNet 프레임워크 개발 시스템의 경우 2,500만 레코드의 자료를 입력하는데 94,000여 초의 시간이 소요되었다. 그러나, 시험운영 서

표 10. 저장엔진과 인덱스 사용 유무에 따른 데이터베이스 검색시간(21억 개 레코드 기준)

index	저장엔진	검색	Searching Time (s)
X	MyISAM	최초검색	346.21
		평균검색시간	348.43
	InnoDB	최초검색	351.47
		평균검색시간	352.37
O	MyISAM	최초검색	3.66
		평균검색시간	0.28
	InnoDB	최초검색	4.24
		평균검색시간	0.28

버의 경우 6,200만 레코드의 자료를 입력하는데 13,000여 초가 소요되었다. 표 3의 결과에서 보듯이 시험운영 데이터베이스 서버는 KMTNet 프레임워크 개발 시스템보다 3배 이상 빠른 I/O 성능을 가지고 있지만 그 외 CPU와 메모리, 각종 디스크 컨트롤러 등으로 인해 더욱 큰 성능 차이가 발생한 것으로 보인다. 2,500만 레코드의 자료를 입력하였을 경우 시험운영 데이터베이스 서버는 프레임워크 개발 시스템에 비해 18배 이상의 입력속도를 보였고, 구축 성능 목표를 만족하였다.

4. 결론 및 토의

우리는 KMTNet 측광 데이터베이스와 측광 파이프라인을 시험운영하기 위한 시스템을 구축하였고, 운영 시 발생할 수 있는 문제점을 찾고 이를 보완하기 위한 실험을 수행하였고 시험운영서버의 구축 성능 목표를 만족하였는지도 분석하였다.

측광 데이터베이스 시스템은 전원 2중화를 통하여 안정성을 확보하였고 측광 파이프라인 시스템은 입출력 성능의 향상에 중점을 두었다. 외부 저장장치는 8 Gbps의 대역폭을 사용하였고, 내부 네트워크 10 Gbps를 사용하는 등 병목현상이 발생할 수 있는 부분을 제거하는데 중점을 두었다. 그로인해 시험운영용 데이터베이스 서버는 KMTNet 파이프라인 개발 시스템보다 18배 이상 빠른 성능을 보였고 측광 파이프라인 서버는 기존 시스템에 비해 5배 이상의 성능을 보였다. 관측소에서 관측 자료를 백업하고 전송할 미디어로 LTO-5 테이프를 선택하였다. 이 미디어는 HDD와 견줄만한 입출력 속도를 보여주었으며 서버 간 호환성 문제도 나타나지 않았을 뿐만 아니라, 전체적인 성능도 KMTNet에서 사용할 목적에 부합된다고 판단되었다.

KMTNet 관측이 시작된 후 5년이 지나면 300 TB 이상의 데이터베이스가 형성이 되는데, 하나의 노드로 검색 서비스를 수행하기는 불가능하기 때문에 시험운영 서버의 데이터베이스 병렬화에 대한 실험을 수행하고 결과를 정리하였다. 결과에 의하면 KMTNet은 NDB Cluster를 이용한 데이터베이스의 병렬화보다 MyISAM이나 InnoDB 저장엔진을 사용하여 관측 영역별로 여러 노드에 데이터베이스를 분산하는 방식이 더 효율적임을 확인하였다. InnoDB 저장엔진은 자료의 입력과 검색이 동시에 가능하여 자료의 입력과 검색이 빈번하게 발생하는 KMTNet 데이터베이스에 적합하지만, 자료 입력시간이 오래 걸리는 문제로 KMTNet 프레임워크 개발에는 속도 위주의 MyISAM 저장엔진을 사용하였다(김동진 등, 2011). 표 8의 CASE 5 실험 결과처럼 입력 자료를 분할한 후 여러 서버에서 job scheduler로 입력하는 방법을 사용한다면 InnoDB의 입력 시간을 1/5로 단축시킬 수 있다. 따라서 KMTNet 데이터베이스 저장엔진을 MyISAM에서 InnoDB로 교체하여 입력과 검색을 동시에 진행할 수 있도록 하였다.

우리는 이 연구를 통해 시험운영 서버를 구축하고 하드웨어와 소프트웨어의 단계별 성능을 점검하였다. 향후에는 이번에 구축한 시험운영 서버를 기반으로 모의 관측 자료를 생성하고 시험운영을 실시하여 측광 파이프라인, 측광 데이터베이스 및 분석모델링 프로그램이 유기적으로 연동하는지 확인하여야 한다. 그리고 시험운영 과정에서 도출된 문제점들을 분석하여 실시간 운영 시 문제가 발생하면 즉각 대처할 수 있도록 하여야 할 것이다.

참고 문헌

김동진, 이충욱, 김승리, 박병곤, 이재우, 2009, KMTNet 자료처리 시스템 설계와 측광데이터베이스 구축, 천문학논총, 24, 83

김동진, 이충욱, 김승리, 박병곤, 2011, 중력렌즈 사건 측광 데이터베이스 및 프레임워크 개발, 천문학 논총, 26, 41

이충욱, 박병곤, 김승리, 이재우 등, 2009, 영상차감법을 이용한 대용량탐색자료처리 파이프라인 개발, 한국천문연구원 기술보고서(No. 20090238)

Bond, I. A., Abe, F., & Dodd, R. J., et al., 2001, Real-Time Difference Imaging Analysis of MOA Galactic Bulge Observations During 2000, MNRAS, 327, 868

Lee, C. -U., Koo, J. -R., Kim, S. -L., Lee, J., Park, B. -G., & Han, C., 2010, Detection of Variable Stars in the Open Cluster M11 Using Difference Image Analysis Pipeline, JASS, 27, 289

Schneider, J., Dedieu, C., Le Sidaner, P., Savalle, R., & Zolotukhin, I., 2011, Defining and Cataloging Exoplanets: the Exoplanet.eu Database, A&A, 532, A79

Udalski, A., Szymanski, M., Soszynski, I., & Poleski, R., 2008, The Optical Gravitational Lensing Experiment.

Final Reductions of the OGLE-III Data, Acta Astronomica, 58, 69

Wolszczan, A. & Frail, D. A., 1992, A Planetary System Around the Millisecond Pulsar PSR1257+12, Nature, 355, 145