대전상관기의 상관결과 고속저장을 위한 데이터아카이브 시스템의 성능시험

Performance Evaluation of Data Archive System for High-Speed Saving of Correlated Result of Daejeon Correlator

노덕규*, 오세진*, 염재환*, 오충식*, 윤영주*, 정진승*, 정동규*

Duk-gyoo Roh*, Se-jin Oh*, Jae-hwan Yeom*, Chung-sik Oh*, Young-joo Yun*, Jin-seung Jung*, Dong-kyu Jung*

요 약

본 논문에서는 대전상관기의 상관결과를 고속으로 저장하기 위한 데이터아카이브 시스템의 성능시험에 대해 기술한다. 대전상관기는 다양한 관측모드를 지원하고 있는데, 이 관측모드 중에서 적분시간에 따라 상관기로부터 출력되는 상관결과의 속도에 영향을 받는다. 대전상관기의 상관결과를 가장 빠른 속도로 출력하는 것은 최대 1.4GB/s를 출력하며, C1 관측모드일 경우 25.6ms의 상관적분인 경우이다. 본 연구에서는 대전상관기의 핵심부분인 VLBI상관서브시스템(VLBI Correlation Subsystem)과 4개의 광케이블로 연결되어 상관결과를 저장하는 데이터아카이브 시스템의 성능시험을 수행하였다. 데이터아카이브 시스템은 본 연구를 위해 2개 회사의 제품을 선정하여 벤치마크테스트(Benchmark Test)를 진행하였다. 본 논문에서는 벤치마크테스트를 위해 개발한 VCS의 상관출력 파일 생성 프로그램과 시험결과에 대해 기술한다.

ABSTRACT

In this paper, we introduce the performance evaluation of data archive system for saving correlation result of Daejeon correlator with high-data rate. Daejeon correlator supports various correlation modes, but the speed of correlation result is affected by correlator according to the integration time in each mode. Maximum data rate of Daejeon correlator is 1.4GB/s in case of C1 mode with 25.6ms integration time. In this research, the performance evaluation of the proposed data archive system is conducted for saving correlation results connected with 4 10GbE optical cable with VCS (VLBI Correlation Subsystem), which is the core system of Daejeon correlator. For the experiments, the data archive system for 2 benders was selected and benchmark test was performed. In this paper, the developed data generation program of VCS correlation result file for benchmark test and evaluation results are described.

Keywords : Daejeon Correlator, Data Archive System, Writing Speed Evaluation

Ⅰ. 서 론

초장기선전파간섭계(Very Long Baseline Interferometry, VLBI)와 같은 전파천문학에서 관측데이터의 데이터처리를 담당하는 장치를 상관기(Correlator)라고 한다[1]. 과거 전파천문학에서 개발한 상관기는 관측데이터의 관측대역폭이 작고 디지털화한 데이터 샘플링 속도가 느려서 상관처리 성능도 그렇게 좋지는 않았다. 그러나 최근 천문학자들의 광대역/고속화 등의 요구사항에 맞추어 관측시스템의 성능이 좋아지면서 관측데이터의 처리를 위한 장치들의 대용량/초고속화가 동반되고 있는 실정이다[1].

본 연구에서는 한국천문연구원과 일본국립천문대가 공동으로 개발한 대전상관기[2][3]의 성능을 만족할 수 있는 차세대 데이터아카이브 시스템의 성능에 대해서 사전조사를 수행하였으며, 대전상관기의 출력과 동등한 가상의 데이터 파일을 설계하고 전송하는 프로그램을 작성하여 사전조사 대상 시스템에 적용하여 성능시험을 수행하고 그 결과에 대해 고찰하고자 한다.

* 한국천문연구원

투고 일자 : 2014. 3. 14 수정완료일자 : 2014. 4. 22
게재확정일자 : 2014. 5. 2

본 논문은 다음과 같이 구성된다. Ⅱ장에서는 대전상관기의 상관출력규격에 대해 간략히 기술하고 Ⅲ장에서는 도입 예정인 데이터아카이브 시스템의 구성과 성능시험 프로그램에 대해 살펴보고, Ⅳ장에서는 성능시험 및 결과에 대해 고찰한 후, 마지막으로 Ⅴ장에서는 본 논문의 결론을 맺는다.

Ⅱ. 대전상관기의 상관출력 규격

대전상관기의 핵심부분인 VLBI상관서브시스템(VLBI Correlation Subsystem, VCS)는 그림 1에 나타낸 구성과 같이 데이터아카이브 시스템과 연결된다. VCS는 16관측국에 대해 최대 8192Mbps, 상관결과당 8192출력채널을 처리하여 최대 상관출력속도는 초당 1.41GB/s의 데이터를 처리할 수 있다[3][4]. VCS에는 서브어레이 모드를 지원하는데, 8+8, 12+4, 16 관측국 모드이다. 본 논문에서는 관측데이터의 상관출력 데이터양이 가장 많은 16관측국 모드에 대해 기술한다. VCS는 32개의 상관블록(Correlation block)으로 구성된다. 각 상관블록의 상관결과는 TCP/IP 프로토콜을 이용하여 독립적으로 GbE로 출력되며, 10 기가비트 스위칭 허브(10GbE SW HUB)는 각 GbE 데이터를 10GbE로 통합하여 4개의 광케이블을 거쳐서 데이터아카이브 시스템으로 전송한다. 그림 2는 1개 상관블록과 데이터아카이브 시스템 사이의 데이터 시퀀스(Sequence)와 데이터 전송구조를 나타낸 것이다.

표 1은 VCS에서 지원하는 대표적인 상관모드를 나타낸 것이다. 한국우주전파관측망(KVN)에 설치된 관측장비중에서 VLBI 관측시에 1024Msps/2bit로 샘플링할 경우 데이터율은 2048Mbps가 된다. 이 데이터율에 따른 데이터를 모두 기록하는 장치가 당시엔 없었기 때문에 디지털 필터를 설치하여 최대 256MHz 대역폭으로 최대 1024Mbps 속도로 필터링을 하였다. 따라서 디지털 필터링과정에서 256MHz 대역폭에 관측 스트림을 몇 개를 설정하는가에 따라 관측모드가 정해지며, 표 1에 나타낸 것과 같이 상관모드도 결정된다.

표 2는 상관결과에서 1개 스트림에 대한 데이터양을 나타낸 것이다. 각 상관블록 1 세트의 상관결과 데이터양은 HEADER 패킷 1개, 자기상관 패킷 16개, 상호상관 패킷 240개로 총 TCP 패킷의 수는 257 패킷으로 구성되며, 총 265,264 byte이다. 그리고 1 패킷당 오버헤더(Overhead)는 Preamble(4 byte), Ether(14 byte), IP HEAD(20 byte), TCP HEAD(20 byte), FCS(4 byte)로서 총 15,934 Byte이다. VCS는 8192Mbps를 지원하기 위해 VSI[1] 규격이 2048Mbps를 지원하기 때문에 4개의 입력포트를 갖고 있다. 따라서 VCS 4개의 포트에 대해 데이터가 출력되므로 1포트당 1개 상관블록의 데이터양(Ethernet 위에서 전송되는 송신양)은 281,198 byte가 된다. 그러므로 4개의 상관출력포트에 총 1,124,792 byte이며, 총 32개 상관블록에 대해서 35,993,344 byte의 데이터가 출력된다. 이 데이터양은 1개 스트림(Stream)에 대한 경우이므로 표 1에 나타낸 C5의 16 스트림인 경우 총 575,893,504 byte의 데이터가 출력된다. 따라서 표 3에 나타낸 것과 같이 적분시간이 1.024s인 경우 562.4MB/s의 상관출력속도가 된다. VCS는 표 3에 나타낸 것과 같이 최대 상관출력속도 1.41GB/s 이상은 대응하지 못하기 때문에 각 스트림 및 적분시간에 따라 데이터양이 제한된다.

1) http://vlbi.org/vsi

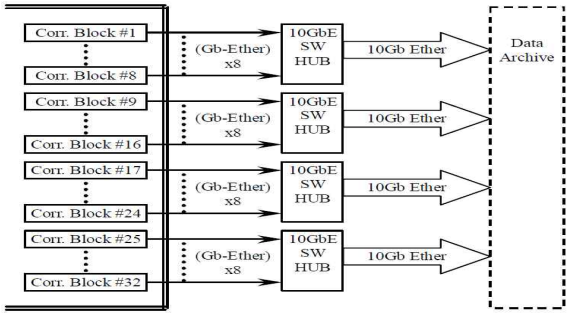

그림 1. VCS와 데이터아카이브 시스템 사이의 10GbE 연결 구성도
Fig .1. 10GbE connection configuration between VCS and data archive system.

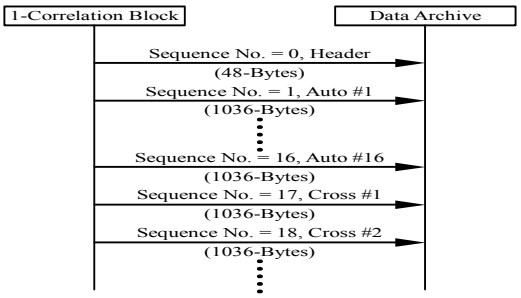

그림 2. 1개 상관블록과 데이터아카이브 시스템 사이의 데이터 시퀀스 및 데이터 전송구조
Fig .2. Data sequence and transmission structure between 1 correlation block and data archive system.

표 1. VCS의 대표 상관처리 모드
Table 1. Representative correlation mode of VCS.

상관모드	대역폭(MHz)	출력IF(Stream)
C1	256	1
C2	128	2
C3	64	4
C4	32	8
C5	16	16

표 2. 상관결과의 데이터양
Table 2. Data quantity of correlation result.

	16-station mode	비고
1개 상관블록당 1set 상관결과크기	265,264-byte	HEADER Packet×1 자기상관 Packet×16 교차상관 Packet×240
TCP Packet 수	257-Packet	
Packet overhead	15,934-byte	1PACKET당 overhead Preamble (4byte) ETHER (14byte) IP HEAD (20byte) TCP HEAD (20byte) FCS (4byte)
1개 상관블록당 1set 데이터양 (이더넷상에서 전송되는 송신량) 1port당	281,198-byte	265,264+15,934Byte (TCP Data를 읽는 것에는 user 쪽에서는 Packet의 Overhead를 생각할 필요없음. TCP Payload Data만 읽음)
Port에 따른 데이터양	1,124,792-byte	281,198 × 4 항상 4-PORT 분량이 출력
32상관블록 전체(1Stream)	35,993,344-byte	1,124,792 × 32

표 3. 적분시간과 스트림수에 따른 VCS 상관 데이터양
Table 3. VCS correlation result data quantity according to the stream number and integration time.

Stream Number	IP 시간 및 데이터양								
	25.6ms	51.2ms	102.4ms	204.8ms	409.6ms	512ms	1.024s	2.048s	10.24s
1	1.41 GB/s	703.0 MB/s	351.50 MB/s	175.7 MB/s	87.8 MB/s	70.30 MB/s	35.15 MB/s	17.57 MB/s	3.52 MB/s
2	2.81 GB/s	1.41 GB/s	703.0 MB/s	351.50 MB/s	175.7 MB/s	140.60 MB/s	70.3 MB/s	35.15 MB/s	7.03 MB/s
4	5.62 GB/s	2.81 GB/s	1.41 GB/s	703.0 MB/s	351.50 MB/s	281.2 MB/s	140.60 MB/s	70.3 MB/s	14.06 MB/s
8	11.25 GB/s	5.62 GB/s	2.81 GB/s	1.41 GB/s	703.0 MB/s	562.4 MB/s	281.20 MB/s	140.60 MB/s	28.12 MB/s
16	22.50 GB/s	11.2 GB/s	5.62 GB/s	2.81 GB/s	1.41 GB/s	1.12 GB/s	562.4 MB/s	281.20 MB/s	56.24 MB/s

표 3은 적분시간과 스트림 수에 따른 VCS 상관결과의 데이터양을 나타낸 것이다. 표 3에서 회색으로 표시된 칼럼은 VCS의 성능을 초과하는 데이터양을 나타낸 것이다. 각 스트림 수에 대해 VCS는 1.41GB/s가 최고속도이며, 16 스트림인 경우 적분시간이 409.6ms일 때 최고속도를 나타낸다. 이는 그만큼 VCS에서 상관결과를 출력할 때 스트림 수와 적분시간의 영향을 받는 것을 알 수 있다.

Ⅲ. 데이터아카이브 시스템의 구성

3.1 HP DL360-P2000

HP사의 최신형 서버인 ProLiant DL560 서버와 SAN 스토리지 P2000로 구성되는 새로운 데이터아카이브 시스템을 제안 받았다. 그림 3의 시스템 구성도에 나타낸 것과 같이 SAN 스토리지 운영을 위한 MDC 서버(DL160) 및 SAN 스위치가 포함되며, 향후 더 많은 저장 공간이 필요한 시점에 P2000만 증설하면 추가적인 용량 증설이 가능하다. 본 연구의 시스템 성능측정을 위해서는 제안된 DL560(표 4)보다는 다소 규모가 작은 DL360 서버와 30TB 급의 P2000 시스템을 데모 장비로 임차하여 기록 속도 측정 실험에 활용하였다.

표 4. HP DL560 서버의 주요 규격
Table 4. Specification of HP DL560 server.

HP ProLiant DL560 G8
- Processor : Intel® Xeon® E5-4610 (2.4GHz/6-core/15MB) × 4EA - Memory : 64GB PC3L-10600R-9 Kit - Network Controller : HP NC523SFP 10Gb 4Port - Network Controller : HP Ethernet 1Gb 2Port 332T Adapter - Expansion Slots : 6 Slots - Storage Controller : HP Smart Array P420i Controller / 512MB FBWC - Hard Disk : HP 300GB 6G SAS 10K 2.5in SC ENT HDD × 2EA - Optica Disk : None - HBA Card : HP 82Q 8Gb Dual Port PCI-e FC HBA × 2EA - Power : HP 1200W CS Platinum Plus Hot Plug × 2EA - Form Factor : 2U Rack form factor

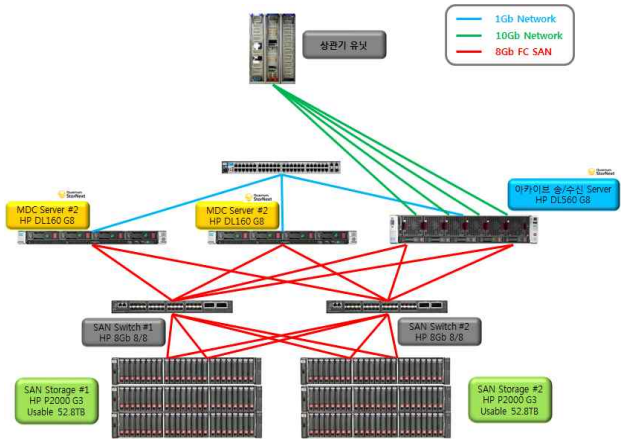

그림 3. HP가 제안한 데이터아카이브 시스템 구성도
Fig .3. Configuration of data archive system proposed by HP.

3.2 Dell PowerEdge R720

한국천문연구원의 한일상관센터에서 DiFX 소프트웨어 상관기 서버로 사용하고 있는 호스트 Virgo는 Dell PowerEdge R720과 레이드(RAID) 컨트롤러를 통해 직접 연결된 2개의 스토리지로 구성되어 있다. 이 서버 시스템에도 최대 4개의 10GbE 포트가 설치되어 있어서, VCS의 상관출력을 직접 받을 수 있으며, 2개의 스토리지를 교호로 사용할 경우, 현재 출력되는 상관출력의 저장 작업과 이전에 저장된 상관출력의 읽기 작업을 분리할 수 있는 장점이 있다. 그림 4에 서버 시스템 및 스토리지의 구성도를 나타내었다. 시스템에 직결된 레이드 디스크는 모두 2개가 있으며, Scratch0는 총 144TB, Scratch1은 총 96TB 용량으로 구성되어 있다. 테스트 당시 Scratch1에서 사용가능한 용량은 약 10TB로 컷지만, 전체적인 기록 속도 측정 목적에는 지장이 없다고 판단하여 실험을 진행하였다.

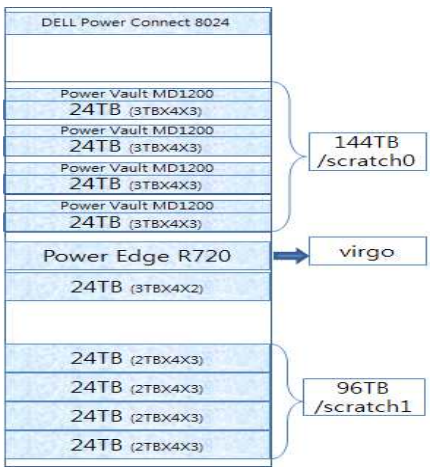

그림 4. 현재 DiFX 소프트웨어 상관기로 사용중인 Dell PowerEdge R720 및 스토리지의 구성
Fig .4. Storage configuration of Dell PowerEdge R720 used in current DiFX software correlator.

데이터아카이브 시스템은 VCS와 4개 회선의 10GbE으로 연결되어, 상관결과 데이터를 수신하고 저장하는 역할을

담당하고 있다. 현재 한일상관센터에서 운용중인 데이터아카이브 시스템은 평균적으로는 최대 1.4GB/s 정도의 고속 데이터를 수신하는 성능을 갖고 있으나, VCS의 하드웨어에서 쏟아지는 Burst data rate을 감당하기에는 다소 부족한 통신 성능을 갖고 있어서, 실제 VCS의 상관출력을 저장할 때 일부 누락되는 TCP/IP 패킷이 발생한다는 점이 판명되었다. 이에 따라 상관처리 시스템의 원활한 운용을 담보하기 위하여, 현 시스템보다 TCP/IP 통신성능이 더 개선된 새로운 데이터아카이브 시스템의 도입이 요청되고 있으며, 최근 하나의 노드에 4개의 10GbE NIC를 설치할 수 있을 정도로 시스템 대역폭이 개선된 서버가 출시됨에 따라, 한일상관센터의 새 데이터아카이브 시스템으로 활용이 가능할지를 검증하는 시험을 진행하였다.

한편, 하나의 노드에서 4회선의 10GbE를 모두 감당할 수 있는 시스템이라는 점은 상관센터 운영의 효율성 제고에 큰 영향을 끼치게 된다. 현재 도입된 시스템은 수년 전에 출시된 것으로 10GbE NIC 1개를 감당하는 노드를 여러 개로 묶어 병렬처리하고 있으나, 최신 노드는 넓어진 대역폭을 활용하여 현재 32 분할로 처리중인 전체 상관출력을 단일 노드에서 한꺼번에 처리할 수 있는 가능성이 있다. 즉, 현재 상관출력을 32개의 중간 파일로 저장한 다음 다시 읽어서 전체의 통합된 데이터로 재편하는 작업을 고쳐서, 중간파일 저장 과정을 생략하고 최종 결과파일을 바로 작성할 수 있는 기반이 갖추어지는 셈이다. 이러한 수정이 이루어지면, 향후 상관처리 작업의 단계를 감축할 수 있어서 상관센터 운영의 효율화를 기대할 수 있다.

Ⅳ. 성능시험 및 결과

4.1 성능시험

본 논문에서는 HP사의 DL360-P2000과 Dell PowerEdge R720 시스템에 대해 동일한 기준으로 기록속도 측정 테스트를 수행하고 그 결과를 분석하였다. 측정에 사용된 프로그램은 VCS의 출력 규격에 맞춰 자체 제작하였으며, 단위 레코드 길이를 지정한 횟수만큼 반복하면서 실제 기록에 소요된 시간을 측정할 수 있도록 고안하였다. 실제 측정에는 최대 32개까지 프로세스 수를 조정하면서 여러 번 측정할 수 있도록 설계한 쉘 스크립트를 활용하였다. 표 5에 이 프로그램의 소스코드 일부분을 나타내었으며, 표 6과 7에 Dell PowerEdge R720의 실험에 사용한 두 개의 쉘 스크립트를 나타내었다. HP P2000용 쉘 스크립트는 파일 저장경로만 다를 뿐 동일하다.

성능측정실험은 쉘 스크립트 run_t1_all을 사용하여 수행하였다. VCS의 1회 적분 상관 데이터 크기를 3600회 반복 기록하면서 매 Write 당 실 소요 시간을 측정하는 실험을 기본 단위로 하여, 프로세스의 개수가 1, 2, 4, 8, 16, 32일 경우에 대해 반복하였다. 또 프로세스 1개를 이용하여 디스크의 용량이 허용하는 한도까지 채우는 실험도 수행하였다.

Ⅱ장의 표 2에서 설명한 것과 같이 VCS 출력의 기본 단위는 265,264 byte이므로, 16 스트림 × 4 포트의 전체에 해당하는 "1회 적분 상관 데이터"의 크기는 16,976,896 byte이다. 본 연구에서는 표 2에 나타낸 것과 같이 TCP Payload Data 크기만을 고려하였다. 즉, 1개 스트림에 대해 265,264×4(port)×32(상관블록)=33,953,792 byte가 되며, 표 3의 C5 상관모드인 경우 33,953,792×16(스트림)=543,260,672 byte가 된다. 성능시험에서는 32개 상관블록에 대해 프로세스의 개수로 시스템이 동작하는지 확인하는 시험을 수행하였다.

표 5. 측정실험에 개발한 프로그램 소스 코드 : t1_write.c.
Table 5. Program source code developed for performance measurement : t1_write.c.

```
// t1_write.c
// compile : $ gcc -Wall -o t1_write t1_write.c -lrt
// 한 적분 만큼의 데이터를 준비하여, 가능한 한 빨리 기록하면서 시간을 측정한다.
// ------------------------------------------------
typedef struct header_t {
  union {
    unsigned int uint[1];
    char HEADMARK[4];              // "HEAD"
  };
  unsigned int TotalSeqNum;          // 0~0xffffffff --> 전체 IP*Stream 일련번호
  unsigned int SeqNum;              // 0 always     --> 이 블럭 내부의 패킷 일련번호
  unsigned char CorrBlockNum;       // 1~32 --> 파일명 out##.dat의 '##'
  unsigned char ArrayMode;           // 0:16, 1:8+8, 2:12+4
  unsigned short int FFTlength;      // 8~256 (i.e. (8~256)x1024 spectral points)
  unsigned int BinningStartChannel; // 0~0x3ff (i.e. binning from 0~(256x1024-1)-th spectral point)
  unsigned  char  BinningFactor[16];   // 0~255,  16 groups (i.e. for each 2^# points)
  unsigned int StreamNum;           // Serial Number of streams, 0~63 = (port#-1)*16 + (stream#-1)
  unsigned  int  IPlength;                  // 1~400, IntegrationTime = (#)x25.6 msec in VCS time scale
  unsigned int IPcount;              // 0~0xffffffff --> 적분구간의 일련번호
} HEADER;
typedef struct corrdata_t {
  union {
    unsigned  int uint[1];        // 1~16  (fixed for sixteen auto-correlations)
    unsigned int SeqNum;     // 17~256 (fixed for 240 cross-correlations, (r,i)x120 baselines)
  };
  unsigned char Xin;         // Station#, 0~15
  unsigned  char  Yin;              // Station#,  0~15 conjugated (dummy(=0) when autocorr)
  unsigned char StreamNum;      // Serial Number of streams, 0~63 = (port#-1)*16 + (stream#-1)
  unsigned   char   DataKind;             //   0:auto, 2:cross_real, 3:cross_imag
  unsigned int ValidSegments; // 0~0xffffffff, # of valid segments
  int data[256];            // 32bits x 256 channels, offset binary
} CORRDATA;
// ------------------------------------------------
//      데이터의      크기는      AryMode(-->Nbaselines), CmpsMode(-->Nsubstreams)에 따라 정해진다.
//   =  {sizeof(HEADER)  +  sizeof(CORRDATA)*(16  +  2*Nbaselines)} * Nsubstreams * Nports
// = {48 + 1,036 * (16 + 2*120)} * 16 * 4
// = 265,264 * 16 * 4
// = 16,976,896 Bytes
//
// 따라서, 265,264 Byte씩 64회 연속기록 또는 16,976,896 Byte를 1회 기록하면, 1회 적분에 상응.
// ------------------------------------------------
```

표 6. 측정실험에 사용한 쉘 스크립트 : run_t1
Table 6. Shell script used in performance measurement : run_t1.

```
#!/bin/bash
#!/bin/bash
# Run several 't1_write' simultaneously.
Nwrite=$1 # How many Integrations to be written
Nrun=$2   # How many processes to be executed simultaneously
LogDir=${Nwrite}x${Nrun}
mkdir -p ${LogDir}
# clean the previous test files
rm -f /scratch1/test/*.dat
for i in `seq 1 1 ${Nrun}`
do
  Outfilename=/scratch1/test/t1_write_${Nwrite}_${i}.dat
  Logfilename=${LogDir}/t1_write_${Nwrite}_${i}.txt
  rm -f ${Outfilename} ${Logfilename}
  echo "./t1_write ${Nwrite} ${Outfilename} > ${Logfilename}
& "
  ./t1_write ${Nwrite} ${Outfilename} > ${Logfilename} &
done
wait
tail -n 6 ${LogDir}/t1_write_${Nwrite}_*.txt
```

표 7. 측정실험에 사용한 쉘 스크립트 : run_t1_all
Table 7. Shell script used in performance measurment :run_t1_all.

```
#!/bin/bash
echo ================================================
df
echo ================================================
for n in 1 2 4 8 16 32
do
  date
  echo Writing 3600 times with ${n}  processes
  rm -f /scratch1/test/*.dat
  ./run_t1 3600 ${n}
  echo ================================================
done
for n in 1        # 2 4 8 16 32
do
  date
  echo Writing until disk full with ${n}  processes
  rm -f /scratch1/test/*.dat
  ./run_t1 0 ${n}
  echo ================================================
done
rm /scratch1/test/*.dat
echo ALL Done.
```

HP DL360 서버와 P2000 시스템에서, 총 3600회 적분에 해당하는 기록을 반복하는 프로세스를 각각 1개, 2개, 4개, 8개, 16개, 32개(32개 상관블록에 해당)를 동시에 실행시키는 6가지 환경에서 기록 실험을 수행하였다. 이 6가지 유형의 실험에서 해당하는 모든 프로세스들의 결과를 종합하여, 총 소요시간 및 각 쓰기 작업(Write operation)의 평균 소요시간을 그림 5에서 비교하였다. 각각의 그림에서 붉은색 실선은 매 쓰기 작업의 시작시각의 변화 추이를 나타내며, 보라색 +표시는 해당 쓰기 작업에 실제 소요된 평균 시간을 나타낸다. P2000은 전반적으로 고른 기록 소요시간을 보여주고 있으나, 프로세스 수가 증가함에 따라 기록에 소요되는 시간이 2가지로 분리되는 특성을 갖고 있음을 알 수 있다. 각각의 실험에서 프로세스 수가 증가함에 따라 개별 프로세스 당 평균 기록률은 1.08, 0.53, 0.25, 0.12, 0.05, 0.02 GiB/s로 감소하고, 모든 프로세스의 기록을 합친 총 기록률은 1.08, 1.07, 1.04, 0.94, 0.79, 0.68 GiB/s로 감소하고 있음을 알 수 있다. 이는 쓰기 작업의 요구 횟수가 증가함에 따라 대기시간이 증가하여 전체적인 성능저하가 발생하고 있다는 것을 의미한다. 이 기록률에 대한 성능은 P2000을 구성하는 하드 디스크의 개별 성능뿐만 아니라 디스크의 개수에도 영향을 받는 것이므로, 실제 도입하게 될 시스템에서는 데모 장비의 하드디스크 구성과 비교하여 더 많은 수의 하드디스크로 구성하는 것이 중요하다는 것을 시사하고 있다.

마지막으로 P2000에서 프로세스 1개로 디스크 용량을 가득 채울 때까지 반복 기록하면서, 각 쓰기 작업의 시작시각 및 소요시간을 그림 6에 나타내었다. 전체적으로 기록한 총 용량은 약 30TB로 약 7.1시간이 소요되었으며, 평균 기록률은 1.09 GiB/s 수준이었다.

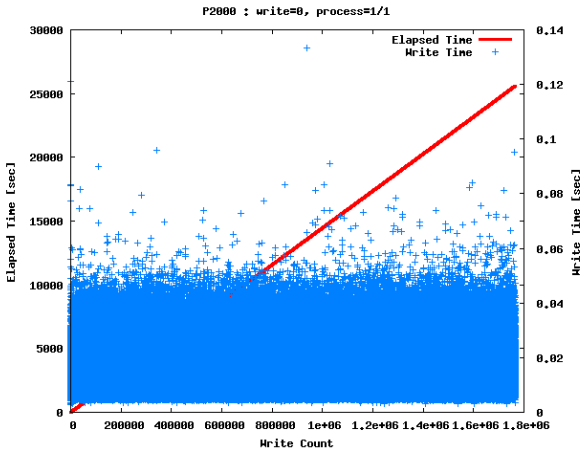

그림 6. P2000에서 프로세스 1개로 디스크 용량을 가득 채울 때까지 기록시 시작시각 및 소요시간
Fig .6. Elapsed time and recording start time until full recording of overall disk capacity with 1 processor at P2000 system.

Dell PowerEdge R720에서도 HP ProLiant DL380-P2000과 마찬가지로, 총 3600회 적분에 해당하는 기록을 반복하는 프로세스를 각각 1개, 2개, 4개, 8개, 16개, 32개를 동시에 실행시키는 6가지 환경에서 기록 실험을 동등하게 수행하였다. 앞에서 수행한 것과 같은 6가지 유형의 실험에서 해당하는 모든 프로세스들의 결과를 종합하여, 총 소요시간 및 각 쓰기 작업의 평균 소요시간을 그림 7에서 비교하였다. 그림 7에서 붉은색 실선은 매 쓰기 작업의 시작시각의 변화 추이를 나타내며, 보라색 +표시는 해당 쓰기 작업에 실제 소요된 평균 시간을 나타낸다. R720에 직결된 레이드 디스크는 개별 쓰기 작업에 대한 기록 소요시간은 시스템의 상태에 따라 3 또는 4가지 유형의 불규칙 반응을 나타내고, 많은 수의 쓰기 작업에 대해서는 통계적으로 고른 기록 소요시간을 보여주고 있으나, 특정 시간대에서는 반응 시간이 급격히 증가하는 특이한 양상을 보인다. 즉 쓰기 작업이 한 번 정체가 되면 해소되기까지 OS가 갖는 부담이 훨씬 크다는 것을 알 수 있다. 한편 프로세스 수가 증가함에 따라 기록에 소요되는 시간이 꾸준히 증가하는 점은 P2000과 같은 양상으로 시스템이 갖는 IO operation의

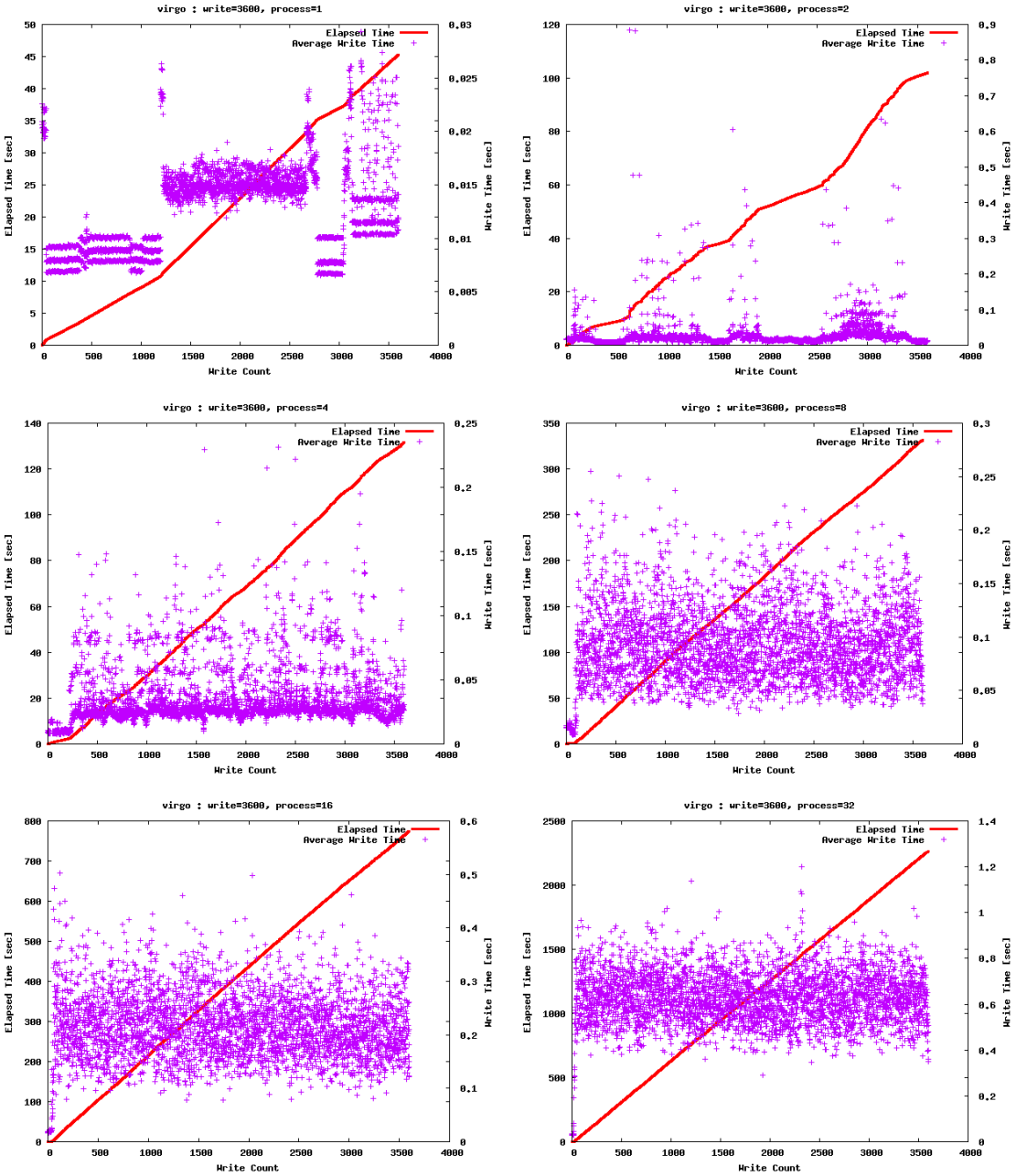

그림 7. R720에서 프로세스 1/2/4/8/16/32개로 각각 3600회 기록시 시작시각 및 평균 소요시간 비교

Fig .7. Comparison of average elapsed time and recording start time for each 3600 times with 1/2/4/8/16/31 process at R720 system.

한계를 의미한다고 판단된다. 각각의 실험에서 프로세스 수가 증가함에 따라 개별 프로세스 당 평균 기록률은 1.25, 0.57, 0.43, 0.17, 0.073, 0.025 GiB/s로 감소하고, 모든 프로세스의 기록을 합친 총 기록률은 1.25, 1.14, 1.74, 1.37, 1.17, 0.80 GiB/s로 변화하고 있음을 알 수 있다. 이는 쓰기 작업의 요구 횟수가 증가함에 따라 대기시간이 증가하여 전체적인 성능저하가 발생하고 있다는 것을 의미하는 가운데, 프로세스 수가 4 또는 8일 때 가장 효율이 좋은 특이한 결과를 보여준다.

이는 서버 본체의 시스템 성능 즉 IO operation 대역폭과 연계된 것으로 해석된다.

마지막으로 R720에서 프로세스 1개로 디스크 용량을 가득 채울 때까지 반복 기록하면서, 각 쓰기 작업의 시작시각 및 소요시간을 그림 8에 나타내었다. 전체적으로 기록한 총 용량은 약 10TB로 약 2.6시간이 소요되었으며, 평균 기록률은 1.04 GiB/s 수준이었다.

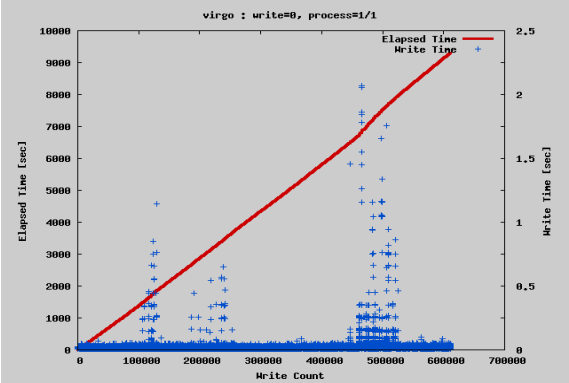

그림 8. R720에서 프로세스 1개로 디스크 용량을 가득 채울 때까지 기록시 시작시각 및 소요시간
Fig .8. Elapsed time and recording start time until full recording of overall disk capacity with 1 processor at R720 system.

4.2 결과고찰

한일상관센터의 대전상관기 시스템에서 제일 마지막에서 VCS의 상관출력을 받아 저장하는 데이터아카이브 시스템은 VCS의 고속 데이터 출력을 손실 없이 저장하는 중요한 역할을 담당하고 있다. 현재 운용중인 데이터아카이브 시스템은 상관기 개발 당시에 도입된 시스템으로서 평균 최대 1.4GB/s 정도의 고속 데이터를 저장하는 성능을 갖고 있으나, VCS의 하드웨어에서 쏟아지는 Burst data rate을 감당하기에는 다소 부족한 통신 성능을 갖고 있어서, 데이터 수신 성능을 향상시킨 새로운 시스템으로 교환하는 것이 요청되고 있다. 이러한 상황에서 10GbE 4 포트를 직접 수신하는 넓은 대역폭의 새로운 서버 및 대용량 스토리지를 테스트하는 목적의 실험을 진행하였다.

첫 번째로, HP사의 최신형 서버인 ProLiant DL560 서버와 SAN 스토리지 P2000로 구성되는 새로운 데이터 아카이브 시스템을 제안 받았으나, 제안 받은 DL560보다는 다소 규모가 작은 DL360 서버와 30TB 급의 P2000 시스템을 데모 장비로 임차하여 사용하였다. 두 번째로는, DiFX 소프트웨어 상관기 서버로 사용하고 있는 Dell PowerEdge R720 시스템 및 직결된 레이드 디스크를 사용하였다.

두 시스템에서 동등한 조건으로 실험을 수행하였으며, 총 3600회 적분에 해당하는 기록을 반복하는 프로세스를 각각 1개, 2개, 4개, 8개, 16개, 32개를 동시에 실행시키는 6가지 환경에서 기록 실험을 수행하고, 마지막으로 프로세스 1개로 디스크 용량을 가득 채울 때까지 반복 기록하는 실험을 수행하였다. 각각의 실험에서 총 소요시간, 각 쓰기 작업의 평균 소요시간을 측정하여 비교, 분석하였다.

P2000의 경우, 전반적으로 고른 기록 소요시간을 보여주고 있으나, 프로세스 수가 증가함에 따라 기록에 소요되는 시간이 두 가지로 분리되는 특성을 갖고 있었다. 각각의 실험에서 프로세스 수가 증가함에 따라 개별 프로세스 당 평균 기록률은 1.08, 0.53, 0.25, 0.12, 0.05, 0.02 GiB/s로 감소하고, 모든 프로세스의 기록을 합친 총 기록률은 1.08, 1.07, 1.04, 0.94, 0.79, 0.68 GiB/s로 감소하였다. 디스크 용량을 모두 채우는 실험에서 전체적으로 기록한 총 용량은 30TB로 약 7.1시간이 소요되었고, 평균 기록률은 1.09 GiB/s 이었다.

R720의 경우, 개별 쓰기 작업에 대한 기록 소요시간은 시스템의 상태에 따라 3 또는 4가지 유형의 불규칙 반응을 나타내고, 많은 수의 쓰기 작업에 대해서는 통계적으로 고른 기록 소요시간을 보여주고 있으나, 특정 시간대에서는 반응 시간이 급격히 증가하는 특이한 양상이 나타났다. 각각의 실험에서 프로세스 수가 증가함에 따라 개별 프로세스 당 평균 기록률은 1.25, 0.57, 0.43, 0.17, 0.073, 0.025 GiB/s로 감소하고, 모든 프로세스의 기록을 합친 총 기록률은 1.25, 1.14, 1.74, 1.37, 1.17, 0.80 GiB/s로 변화하였다. 특이할 점은 쓰기 작업의 요구 횟수가 증가함에 따라 대기시간이 증가하여 전체적인 성능저하가 발생하지만, 프로세스 수가 4 또는 8일 때 가장 효율이 좋은 결과를 보여주는 점으로, 서버 본체의 시스템 성능 즉 IO operation 대역폭과 연계된 것으로 판단된다. 디스크 용량을 모두 채우는 실험에서 전체적으로 기록한 총 용량은 10TB로 약 2.6시간이 소요되었으며, 평균 기록률은 1.04 GiB/s 이었다.

V. 결 론

본 논문에서는 대전상관기의 상관결과를 고속으로 저장하기 위한 새로운 데이터아카이브 시스템을 도입하기 위한 사전조사연구로서 제안된 시스템의 성능시험을 수행하였다. 본 논문에서는 VCS 상관결과와 같은 규격의 파일을 생성하고 기록하는 프로그램을 개발하였고, 제안된 데이터아카이브 시스템을 대상으로 기록성능시험을 수행하였으며, 그 결과를 요약하면 다음과 같다.

DL360-P2000은 데모 장비 구성상의 한계 성능이 약 1 GiB/s이나 개별 쓰기 작업은 고른 반응을 나타내며, 프로세스의 수가 증가함에 따라 반비례하여 평균 기록률이 감소하여 일반적인 성능 저하를 나타내었다. R720-레이드는 개별 쓰기 작업이 불규칙적인 반응을 가지되 통계적으로는 고른 반응을 나타내지만, 일정 수준 이상의 쓰기 작업이 쌓이면 급격한 성능 저하를 나타내는 특성이 있었다. 평균적인 기록률은 1 GiB/s 수준이었지만, 프로세스 수가 4 또는 8인 경우에 평균 기록률이 1.74 또는 1.37 GiB/s의 피크 값을 보였다. 두 시스템 모두 전체적으로 유사한 1 GiB/s 급의 기록 성능을 나타내어 기대하였던 1.4 GiB/s의 성능에는 미치지 못하였지만, 개별 디스크의 회전 속도, 소속 디스크의 개수 등 여러 가지 고려할 만한 튜닝 과정을 거치면 목표 성능을 달성할 수 있을 것으로 기대된다.

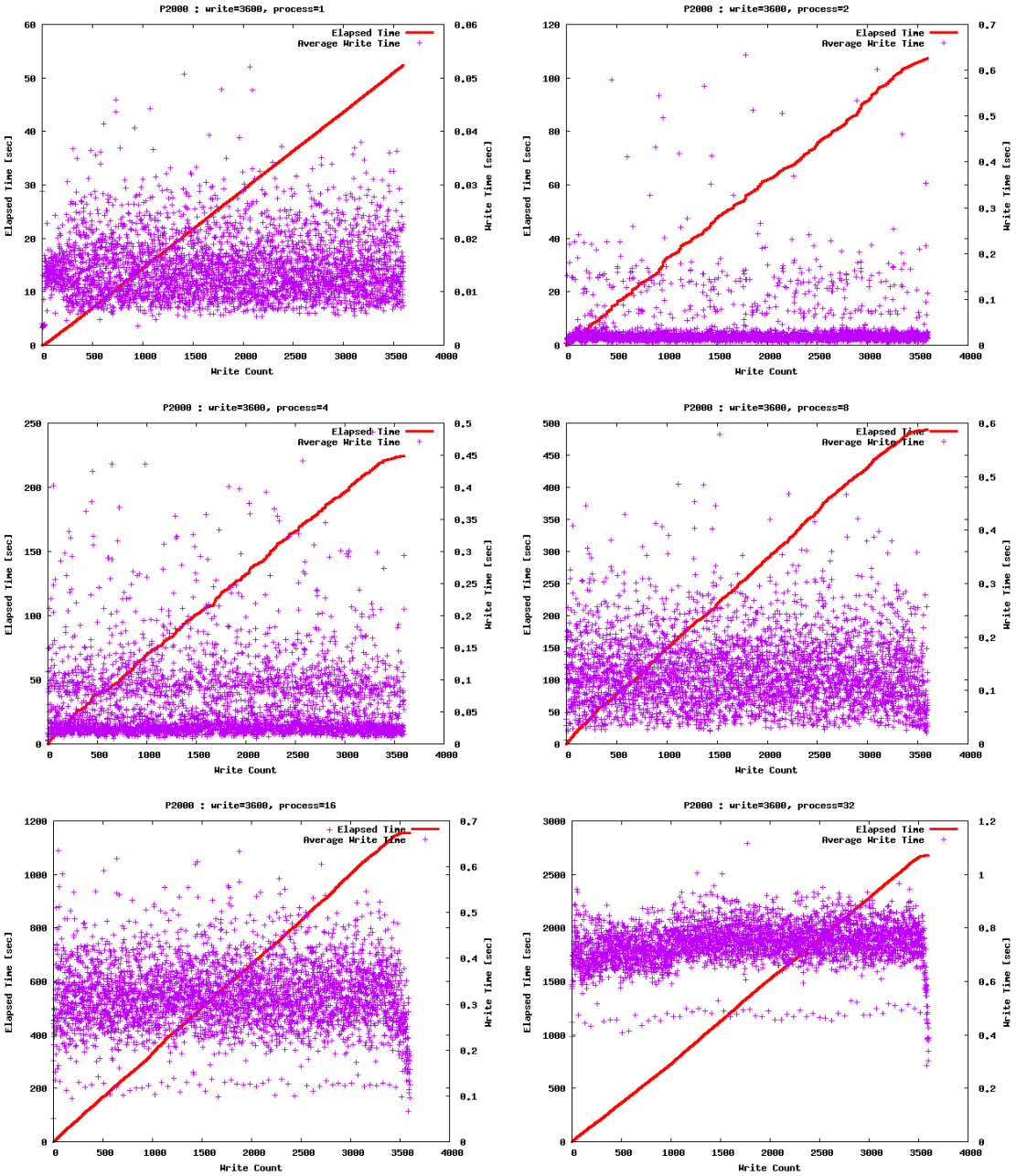

그림 5. P2000에서 프로세스 1/2/4/8/16/32개로 각각 3600회 기록시 시작시각 및 평균 소요시간 비교

Fig .5. Comparison of average elapsed time and recording start time for each 3600 times with 1/2/4/8/16/31 process at P2000 system.

참 고 문 헌

[1] 노덕규, 오세진, 염재환 외 15명, “2008년도 한일공동 VLBI상관기 및 수신기 개발 결과보고서,” 한국천문연구원, pp. 3-100, 2008.

[2] 오세진, 노덕규, 염재환 외 5명, “VLBI상관서브시스템 시작품의 개발에 관한 연구,” 천문학논총 Vol. 24, No. 4, pp. 65-81, 2009.

[3] 오세진, 노덕규, 염재환 외 6명, “VLBI상관서브시스템 본제품의 제작현장 성능시험,” 신호처리시스템학회 논문지 Vol. 12, No. 4, pp. 322-331, 2011.

[4] 오세진, 칸야 유키토시, 노덕규 외 5명, “상관결과 분석을 위한 파일 시스템 설계 및 소프트웨어 개발,” 신호처리시스템학회 논문지 Vol. 14, No. 3, pp. 181-190, 2013.

노 덕 규 (Duk-gyoo Roh)
正會員
1985년 2월 서울대 천문학과(이학사)
1994년 8월 동경대 천문학과(이학석사)
1997년 8월 동경대 천문학과(박사수료)

1985년 4월 ~ 현재 한국천문연구원 책임연구원
2005년 11월 ~ 2009년 3월 한국천문연구원 그룹장
※주관심분야 : 전파천문, VLBI상관기 개발

오 세 진 (Se-jin Oh)
正會員
1996년 2월 영남대 전자공학과(학사)
1998년 2월 영남대 전자공학과(석사)
2002년 2월 영남대 전자공학과(박사)

2001년 9월 ~ 2002년 12월 대구과학대학 교수
2010년 6월 ~ 2011년 5월 한국천문연구원 상관기그룹장
2002년 12월 ~ 현재 한국천문연구원 선임연구원
※주관심분야 : 디지털신호처리, VLBI상관기 개발,
천문관측기기개발

염 재 환 (Jae-hwan Yeom)

2005년 8월 한양대 정밀기계공(석사)
2005년~현재 한국천문연구원 선임연구원

※주관심분야 : 디지털신호처리, VLBI상관기 개발

오 충 식 (Chung-sik Oh)
2002년 2월 서울대 천문학과(이학사)
2006년 3월 동경대 천문학과(이학석사)
2009년3월동경대 천문학과(이학박사)
2009년 4월-2010년 11월 한국천문연구원
박사후연수원

2010년 12월 - 현재 한국천문연구원 선임연구원
※주관심분야 : 전파천문, Astrometry, VLBI상관처리

윤 영 주 (Young-joo Yun)
1999년 2월 서울대 천문학과(이학사)
2001년 2월 서울대 천문학과(이학석사)
2011년2월서울대 천문학과(이학박사)
2011년 3월 ~ 현재 한국천문연구원
선임 연구원

※주관심분야 : 전파천문, VLBI상관기 소프트웨어 개발

정 진 승 (Jin-seung Jung)
正會員
2008년 2월 경남대 전자공학과(공학사)
2010년 2월 경남대 전자공학과(공학석사)
2010년 8월 ~ 현재 한국천문연구원 연구원)

※주관심분야 : 디지털신호처리, FPGA 설계, 천문관측기기
개발

정 동 규 (Dong-kyu Jung)

2004년 8월 충남대 천문학과(이학사)
2006년 8월 충남대 천문학과(석사수료)
2012년 1월 ~ 현재 한국천문연구원 연구원

※주관심분야 : VLBI상관처리, 천문관측기기 개발