

Update on SDSS-II DAQ Upgrade

Kurt Biery

06-Dec-2005

CD Activity Coordination Meeting

Outline of this talk

- Review of changes requested in the DAQ upgrade
- Report on changes that have been implemented
- Description of work that remains to be done

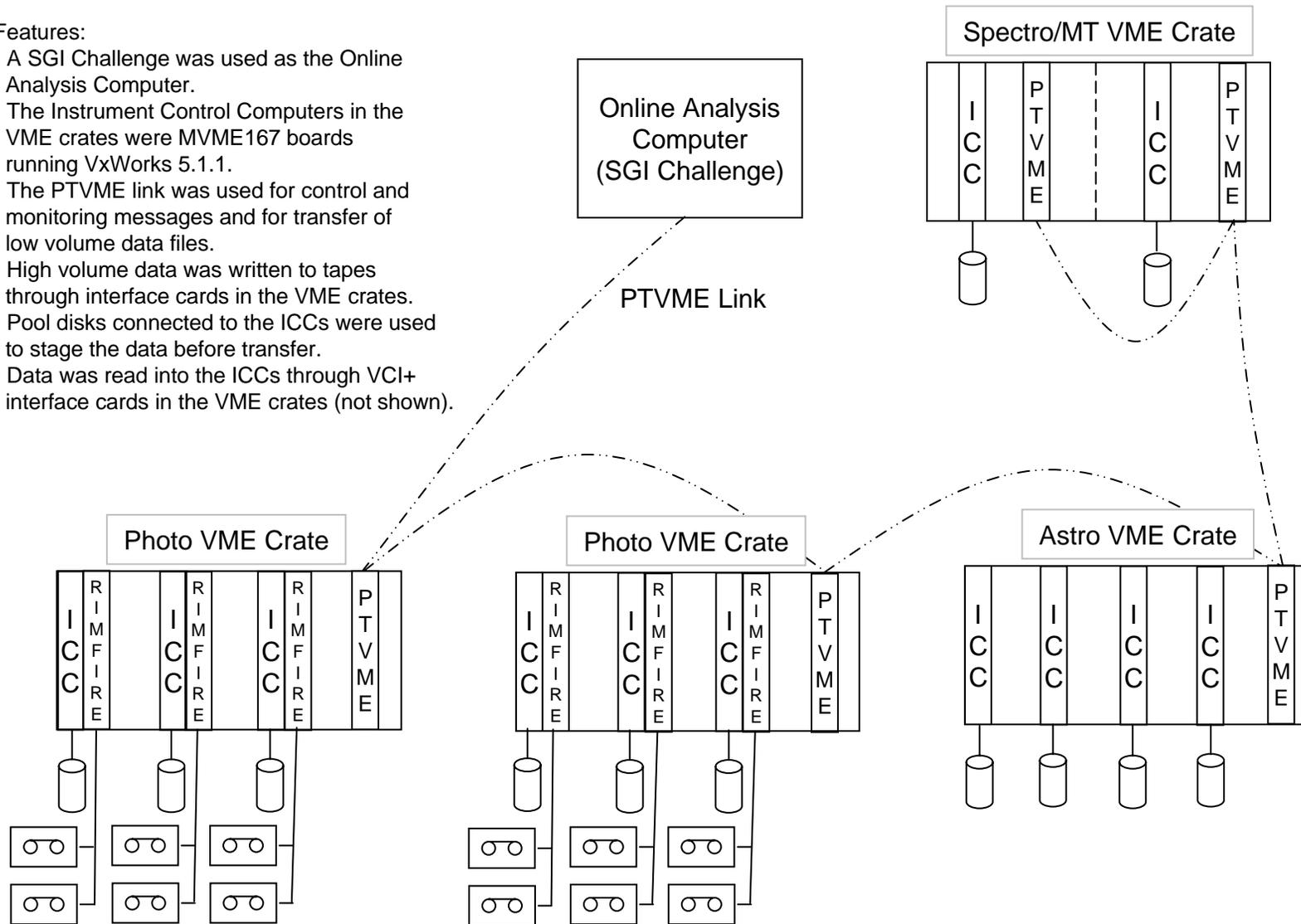
Motivation for the upgrade

- Improve the stability and reliability of the system
- Eliminate dependence on obsolete components
- Reduce operating costs

SDSS-I DAQ System

Features:

- A SGI Challenge was used as the Online Analysis Computer.
- The Instrument Control Computers in the VME crates were MVME167 boards running VxWorks 5.1.1.
- The PTVME link was used for control and monitoring messages and for transfer of low volume data files.
- High volume data was written to tapes through interface cards in the VME crates.
- Pool disks connected to the ICCs were used to stage the data before transfer.
- Data was read into the ICCs through VCI+ interface cards in the VME crates (not shown).



Functional Specification - Hardware Replacements

- Replace the SGI host computer with 1+ Linux PCs
- Replace the MVME167 single board computers with MVME5500s (VxWorks 5.1.1 → 5.5.1)
- Replace the PTVME link with Gigabit Ethernet connections between each MVME5500 and the Linux host(s)
- Remove the tape drives attached to the VME crates and replace that functionality with FTP transfers to Linux disk

Functional Specification - Design Goals

- As much as practical, the existing operator programs will run without modification.
- The weakest elements of the existing DAQ system – the PTLINK, the SGI IRIX OAC, the MVME167 ICCs, and the DLT4000 tape drives - will be eliminated in the upgrade.
- As much as practical, the existing front end readout and analysis code will remain unmodified.

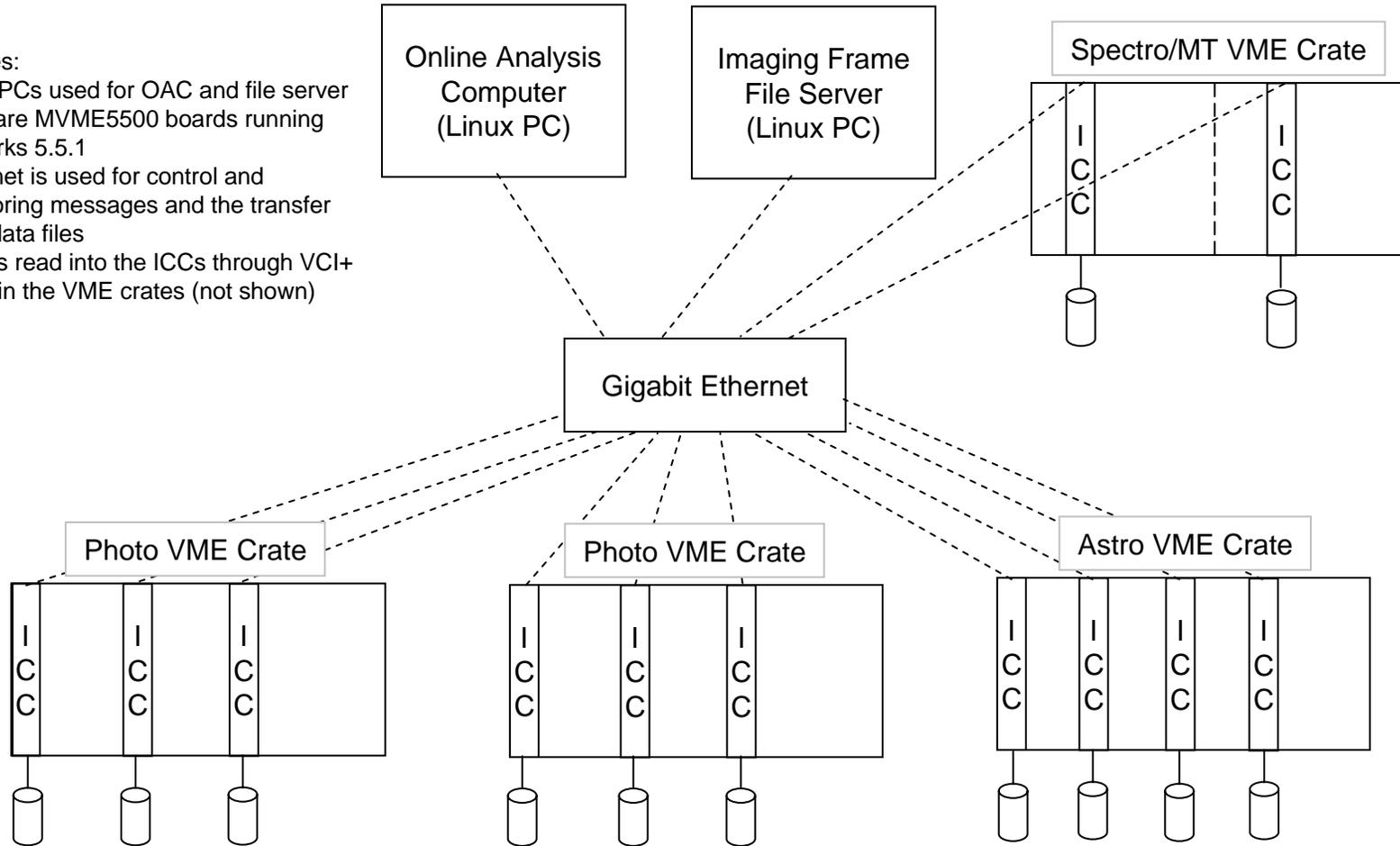
Additional Enhancements

- Addition of FITS checksums when the data is written to the ICC pool disk (VME) [making use of the extra CPU power available on the 5500]
- Improvements in system stability at the application level [adding to the increased stability of the new hardware]
- Network transfer of data to FNAL and Princeton (handled by data processing group). [completely removing the use of tapes] Photo frames are compressed and an MD5sum calculated before transfer.

SDSS-II DAQ System

Features:

- Linux PCs used for OAC and file server
- ICCs are MVME5500 boards running VxWorks 5.5.1
- Ethernet is used for control and monitoring messages and the transfer of all data files
- Data is read into the ICCs through VCI+ cards in the VME crates (not shown)



Replacement of IRIX Host

- The existing IRIX code was ported to Linux.
- A dedicated file server is now used to stage imaging frames (previously written to tape). This server has sufficient disk space to hold ~20 nights of compressed data.
 - The ICC boards transfer the data to this server in real time using FTP
 - Data is copied to FNAL and the supernova analysis cluster at APO from this server
- Hot spares for both the online analysis computer and the file server are online at APO.

Replacement of VxWorks Processors

- A number of issues had to be worked through in porting the existing code from the 167s to the 5500s. Some of the more painful ones:
 - Mapping of full VME address space – required some additional internal bus mapping for the 5500 from Ron
 - DMA transfers on the 5500 must have the three least-significant bits identical in the local and VME addresses – required intermediate copies of the data to valid temporary addresses
 - Lost interrupts from the VCI+ to the 5500 – much worse in the spectro sub-system where one 5500 reads out two VCI+ cards – a work-around is in place, and we don't yet have an understanding of the root cause

PTVME Replacement

- The choice was made to update the existing code at the level of individual reads and writes between the host computer (OAC) and the VME processors (ICC).
- The new system makes use of a server process on each ICC that listens for read and write requests. For a read request, it sends the requested data back to the host, and for a write request, it copies the specified data to the requested address on the ICC.
- This choice allowed us to avoid changes to the higher-level code except in a handful of places that needed data to be byte-swapped.
- As part of minimizing changes at the operator level, the existing system of selecting the target ICC within the operator programs using aliases and crate/board numbers was retained even though the addressing of each ICC changed from PTVME crate and board number to host name in the low-level code.

FITS Checksum

- DATASUM and CHECKSUM keywords were added to the FITS header.
- The DATASUM value is the 32-bit ones complement sum of the data in the file.
- The CHECKSUM value is chosen so that the 32-bit ones complement sum of the entire file equals 0xFFFFFFFF.
- As of 21-Nov-2005, all files have the checksum – photo frames, photo gangs, astro gangs, spectro frames, and pt frames.

System Stability

- Daily reboots of the system are no longer needed.
- Use of system resources has been made more robust (e.g. IPC services no longer left orphaned).
- The frequency of system hangs has been reduced by the elimination of the PTVME link, but there are occasional hangs of individual ICCs that need to be investigated.

Milestones

- Successfully read out internally generated and simulated data through the 5500 to pool disk and the scrolling display at the WH8 test stand on 22-Jul-2005. [The goal was to have this working by early June.]
- The first spectro science data was taken during the night of 29-Aug-2005. [The observing schedule called for 27-Aug-2005 to be the first night of the dark run.]
- The first supernova science runs were taken during the night of 02-Sep-2005 (included a 6.25 hour run). [Supernova observing was scheduled to start 01-Sep-2005.]

Schedule Notes

Challenges:

- We lost Dinker Charak's VxWorks expertise early in the project when he returned to India.
- The hiring of Tom Nicinski for this work did not happen as hoped.
- Steve Foulkes was hired late in the project, too late to help.

Extra help:

- Fritz Stauffer from APO spent two weeks here helping with VxWorks kernel, SCSI disk, and FTP data transfer issues. This is in addition to the support and development that Fritz contributed while at APO.
- Ron Rechenmacher and Don Holmgren
- Margaret Votava, Eric Neilsen, and Craig Loomis
- Lots of others in the computing division and the EAG group

System Performance

- Transfer rates for status queries, copies of data from VME to Linux, etc. (operations that use the TCP/ethernet replacement for PTVME) are in the range of 12 – 15 MByte/sec.
- Five imaging files (5.5 MByte each) are created every 36 seconds by each of the six photo ICCs. The FTP upload of these 30 files from the ICCs to the Linux file server takes 5-6 seconds.
- The readout and processing of each line of CCD data in the ICC (through the VCI+ card) must take less than ~26 msec. After the addition of the checksum calculation, this processing takes ~2 msec.
- The transfer of imaging data to FNAL is done by late morning or early afternoon after each night of photometric data taking. Additional bandwidth on the microwave link to APO was recently purchased to increase the rate to 17.8 Mbit/sec from 12.8 Mbit/sec.

Remaining Work (within my scope)

- Create a monitor process that will run on each ICC to check on the health of ICC processes. Problems will be reported to murmur and picked up by high-level monitors from there. [40-60 hours]
- Update the PTVME replacement code so that it reuses TCP connections and handles clients on different Linux nodes [40-100 hours]
- Move the remaining build that needs to be done on IRIX to Linux (video card replacement for Vigra) [8-12 hours]
- Code cleanup and review of critical code changes [16-24 hours]
- Update the documentation to reflect the new system [40-80 hours?]
- Support [2-10 hours/week?]
 - Assist observers in recovering from problems
 - Assist APO staff in cases of hardware failures

Remaining Work 2 (my scope)

- Verify that binned scanning works and implement any needed fixes
- Understand and fix the occasional ICC board hangs
- Make any needed code changes as various servers (e.g. NTP) move to their final host PCs

- Problem Reports that are currently assigned to me:
 - astrom chips need better sky-subtraction
 - No timestamp/sequence number on astroline measurements

Remaining Work 3 (my scope)

The following items are currently considered to have low priority, but must be completed to improve system reliability and provide a viable and certified test stand at Fermilab for future troubleshooting and support:

- Validate a standard VME crate layout for each sub-system type to serve as a guide for the crates at APO
- Determine the root cause of the lost VCI+ interrupts
- Understand the source of the simulated data mismatches at the WH8 test stand
- WH8 test stand fixes (e.g. get a third pool disk working)

Other items identified by the collaboration...