DAQ: current status & plans

Alexey Filin,
OKA, IHEP, Protvino, September 29, 2011

Prologue

- Neither presentations nor notes about status and architecture of OKA DAQ were done before. The presentation starts to fill the gap
- The presentation is dedicated to current status of OKA DAQ and future plans. It is intended for familiar audience
- Review of OKA DAQ, its architecture and properties is to be done in future
- Some statements are based on measurements to be described in coming reports

Infrastructure (1)

- New dedicated server room with proper cooling, power supply & stand was arranged by V.Lishin
- Cluster computing nodes, file server, oka04, network switches, old pc stand, spares were moved to the new server room. Network cables were made (with help of S.Kholodenko) & laid for front-end & operator DAQ pc's
- Cluster node okaf002 is back. It was required to move system from IDE to SATA disk (there is no driver in SLC4 kernel for IDE controller of okaf002 motherboard!) and rebuild Linux initrd with SATA modules

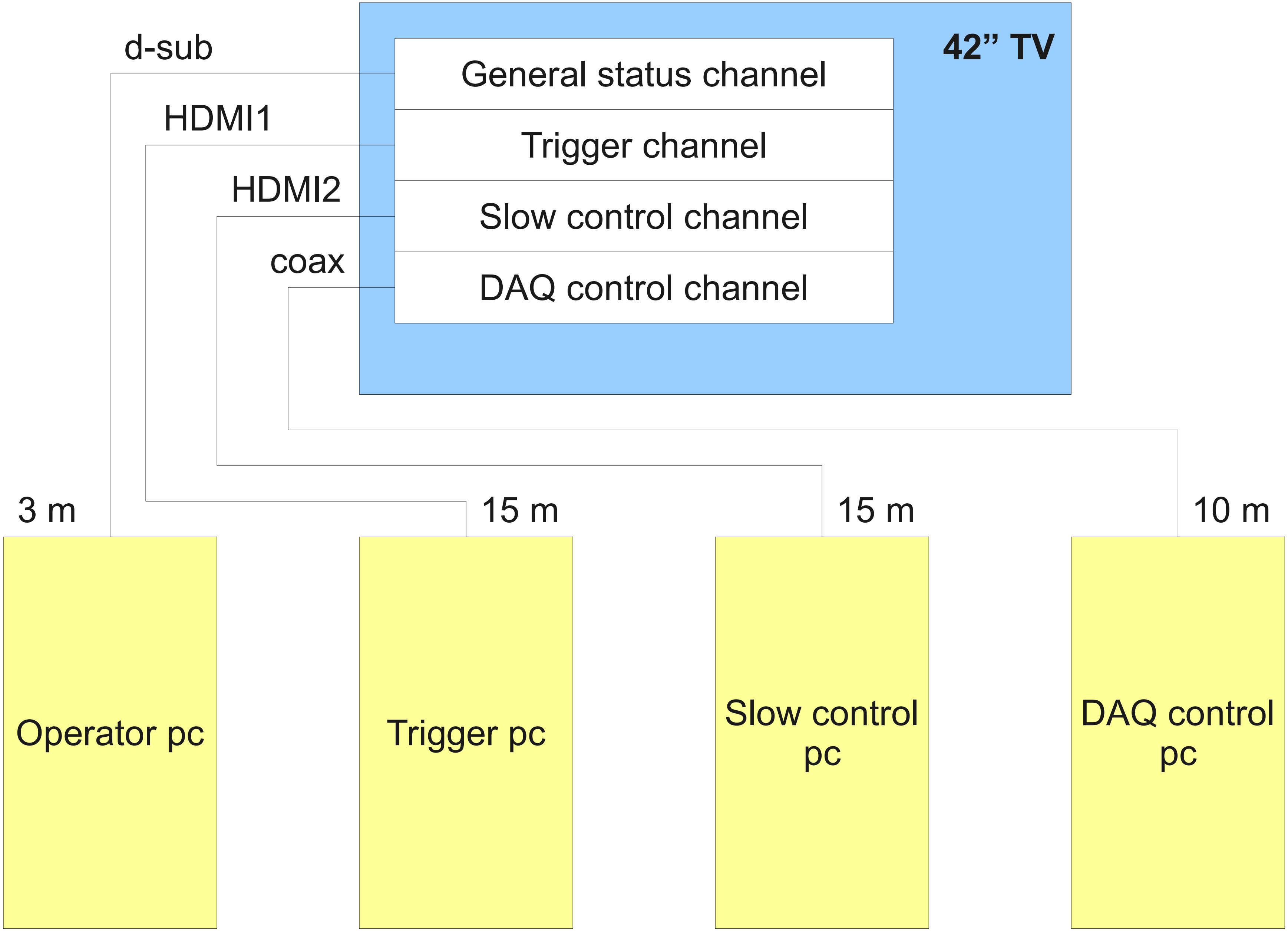
Infrastructure (2)

- New KVM switch for up to 16 pc's was bought & installed. 10 kvm cables were bough only for existing pc's (cause cable is expensive ~$30). CRT monitor can't be used with the switch
- New 4 cluster pc's handed by directorate are installed. It was required to buy UPSes & network interfaces for the pc's. One of new pc's is unstable, its system hangs => pc is under investigation
- Five 3TB hard disks were bought to extend cluster storage for data of coming setup run & generated MC data (thanks to S.Akimenko)

Infrastructure (3)

- Big 42” LCD TV was bought for tea room on second floor to provide real time information about DAQ, trigger and slow control status. It is required to prepare shelf and buy two 15 meter HDMI cables. Four TV channels can be set up:
  - Trigger channel with HDMI1 signal from trigger pc
  - Slow control with HDMI2 signal from slow control pc
  - General info with d-sub signal from operator pc
  - Signal by coaxial cable from another source
- Three old CRT monitors on shelf with old pc are prepared to provide required information for DAQ operators in control room

Information panel



42” TV

General status channel

Trigger channel

Slow control channel

DAQ control channel

d-sub

HDMI1

HDMI2

coax

3 m

15 m

15 m

10 m

Operator pc

Trigger pc

Slow control pc

DAQ control pc

Infrastructure (4)

- 4 cluster computing nodes were upgraded last year with very limited budget (Athlon x4 3GHz has no L3 cache). MC data are generated ~5 times slower on the nodes than on DAQ front-end hosts (Core2 Duo 3GHz, report is to be done by V.Kurshetsov at the next meeting)
- To use infrastructure effectively the nodes should be upgraded (cpu+mb+mem)
- Existing 24 port Gb switches can be used to connect 8 new nodes. Cluster can be extended up to 16 computing nodes (8 is used now)

Software (1)

- Amount of used pc's required to automate pc management (installation, maintenance) => required sw services were installed & tuned:
  - DHCP service (local IP address assignment)
  - Local NFS server with installation repository
  - Kickstart files (automatic installation)
  - PXE booting (booting without local boot image)
  - Network gateway to IHEP & global network
  - Name service in local OKA network is provided with files, in extern network it is provided with IHEP DNS server. Local DNS server can be setup when required.

Software (2)

- LE85 header is back in data by request for analysis of front-end and read-out electronics operation
- Memory leak in MISS-USB driver was detected and removed, so front-end pc's should be stable. Please check crates, make DAQ clean-up before front-end pc rebooting!
- Increased number of incompatibility problems required system upgrade. Fedora 14 is installed on DAQ & cluster pc's. DATE, hw drivers, monitor & calibration sw should be recompiled and rebuilt for Fedora 14 x86_64

Software (3)

- Crate Interface Library:
  - Fixed bug in PCI-Qbus code. The library can be used to access MISS with PCI-Qbus interface. (request from Romanovsky)
  - Added python modules to operate with CAMAC custom made electronics used to generate and commutate physics, LED & PED trigger strobes and accelerator gate (GAMS legacy)
- A set of scripts to automate trigger strobe generation and commutation was developed and is used actively in tests of LE71

Software (4)

- Data decoding library (ddl) was developed to provide official data decoding sw:
  - Modified Kurshetsov database is used to configure decoding, calibration, alignment
  - Data of autumn 2010 and spring 2011 runs are provided with added tables
- Decoding error monitor (dem) was developed with ddl to provide statistics for:
  - 50 types of decoding errors in each front-end & read-out module, equipment, DAQ host
  - Data size in each front-end & read-out module, equipment, DAQ host

Software (5)

- Logbook is under development. Database schema is basically ready (34 tables), scripts to operate with it should be developed
- On-line filter (software trigger) can be implemented when cluster nodes will be tuned as DAQ event builders:
  - Ddl can be easily reconfigured by shift crew
  - A script to check configuration with sequential number read in crates will be run during DAQ start
  - V.Kurshetsov promised to provide correct alignment/calibration on-line + sw filter
  - Cluster and DAQ nodes should run the same OS

Software (6)

- GlusterFS demonstrated its value in practice for 4 years. All real and MC data are kept by 14.6 TB raw disk space on 7 servers. Some drawbacks force to try other variants:
  - Event builder write buffering is a must. Without it direct write performance decreases in ~100 times
  - GlusterFS data duplicating means DAQ event builder output stream = 2x input stream
  - There is no fail over for open file if one of used glfs servers is down => client crashes. No direct write of real data from event builders is available 24/7/365
  - Glfs2 metadata server is a Single Point of Failure
- Lustre cluster fs is under investigation

Electronics (1)

- A set of serious problems in LE71 operation were found by V.Kurshetsov and reported to M.Soldatov after spring run in 2011
- A dedicated test to exhibit LE71 problems was developed, results were reported on a meeting in OEA. The test is used to verify new firmware versions from M.Soldatov. The number of problems on the test decreased considerably
- A new firmware version with synchronized units (to provide identical firmware for different mod #) is developed and promised to be ready soon. All LE71 should be flashed before autumn run

Electronics (2)

- Move from LE75+PCI7200 to LE94+USB2 read-out proved itself:
  - Disappeared transfer errors
  - Transfer speed increased from 6 MB/s to 18 MB/s
  - Developed missusb driver replaced vendor binary driver built for limited number of kernels versions
  - PCI7200 costs about $250 each, USB2 controllers are built into motherboards, discrete USB2 controller costs about $50
  - Custom made flat cables for PCI7200 were replaced with cheap commodity USB2 cables

Electronics (3)

- New read-out module (autonomous controller) to replace LE85+LE94 was developed in OEA:
  - Direct read-out of MISS crate with USB. No converter module (LE94) and flat cable are required
  - Replaceable USB2 interface board in the module provides way to upgrade to USB3 without need to redesing and manufacture the module entirely
  - Replaceable memory modules provide way to extend size of spill buffer when required
  - Built-in dead time measurement
  - One module should be provided for tests next run. 14 modules are planned to be produced to spring run next year to upgrade OKA DAQ

Electronics (4)

- 14 USB cables by HAMA with double shielding and gold plated contacts were bought and tested successfully with LE94:
  - eight 5 meter cables
  - six 3 meter cables
  - two 7.5 meter cables (USB standard declares 5 meter max length!)
- Motherboard of front-end pc's contains two USB controllers with 40 MB/s aggregate bandwith each => up to 4 MISS USB read-out modules (20 MB/s measured bandwidth) can be used without pc upgrade.

Future (1)

- Server room cooling and power supply should be estimated to extend cluster with 8 extra nodes (if sw filter will require extra cpu's)
- Bonding provides way to increase network bandwidth without expensive transition to 10G:
  - 24 port 1Gb switch is to be bought (2 if required)
  - Third 1Gb NIC for each node is to be bought (or 2/4 port NIC should be installed instead of 1 Gb NIC's)
- Back up storage (LTO5 streamer, 1.5 TB tape, LTFS support) was built into next year budget and should provide reserve data storage and repairing for 20-30 years

Future (2)

- USB3 interface should provide bandwidth up to 300 MB/s (half of theoretic bandwidth 600 MB/s) per read-out module. USB3 interface boards are to be manufactured only for crates.
- MISS has no future (CMOS replaced ECL). Its bus reached bandwidth limit (100 ns per word) and holds DAQ dead time. EuroMISS is based on CMOS. One EuroMISS crate is to be provided soon for tests with OKA DAQ. Bus bandwidth remains the problem. Bus should be replaced with serial links (e.g. USB2) to decrease DAQ dead time.

Epilogue

- UNIX Rule of Extensibility: Design for the future, because it will be here sooner than you think (http://en.wikipedia.org/wiki/Unix_philosophy)
- The presentation can be get by http://www.oka.ihep.ru/Members/filin/files/daq_2011sep29.pdf/download
- To be continued...