The ATLAS Read-Out System
Performance with first data and perspective for the future

A. Misiejuk\textsuperscript{1}, G. Crone\textsuperscript{2}, D. Della Volpe\textsuperscript{3}, B. Gorini\textsuperscript{4}, B. Green\textsuperscript{1}, M. Joos\textsuperscript{4}, G. Kieft\textsuperscript{5}, K. Kordas\textsuperscript{8}, A. Kugel\textsuperscript{6}, N. Schroer\textsuperscript{6}, P. Teixeira-Dias\textsuperscript{1}, L. Tremblet\textsuperscript{4}, J. Vermeulen\textsuperscript{5}, F. Wickens\textsuperscript{7}, P. Werner\textsuperscript{4}

\textsuperscript{1}Royal Holloway University of London, \textsuperscript{2}University College London, \textsuperscript{3}Universita & INFN, Napoli, \textsuperscript{4}CERN, \textsuperscript{5}Nikhef, Amsterdam, \textsuperscript{6}Ruprecht-Karls-Universitaet Heidelberg, \textsuperscript{7}Rutherford Appleton Laboratory, \textsuperscript{8}University Bern
System Overview

- **Level 1 Trigger**: 100 kHz
- **L2-Trigger**: 3 kHz
- **Event Filter**: 200 Hz / 300MB/s
- **Event data**
- **Optional network (UDP)**
- **L2 requests @ 20 kHz based on UDP for scaling reasons**
- **ROBIN**: ~600 ROBINs
- **ROS - PC**: ~1600 optical links (1 – 15 per ROS) @ 160 MB/s
- **Fragment rate**: 100 kHz
- **EB requests @ 3 kHz**: Based on TCP/IP due to large message size
- **~150 PCs**
- **~150 PCs**
- **~600 ROBINs**
- **Event Filter**: L2 farm
- **Event Filter**: EB farm
- **4-port NIC**
- **ROD**: ~150 PCs
- **Front-end electronics**
- **ATLAS Detector**: L2-network
- **L2-network**
- **L2-network**

---

Markus Joos – The ATLAS Read-Out System: performance with first data and perspective for future
Building blocks

The ReadOut System (ROS) PC
- Houses 1 to 5 ROBIN cards (typically 4 cards)
- Configures and controls the ROBINs
- Reads data from the ROBINs and provides it to the Second-Level Trigger and to the Event Builder
- Receives clear requests for event fragments and forwards them to the ROBINs
- Interfaces to the operational and physics monitoring systems

The ROBIN PCI card
- Receives event fragments from sub-detector specific front-end electronics (RODs) via 3 optical links
- Buffers events during the decision latency of the Second-Level Trigger and the time required for building of events accepted by L2
- The optical link is based on the S-LINK interface: 32 bit @ 40MHz = 160 MB/s
Hardware Components

- Single Intel Xeon @ 3.4 GHz (Irwindale)
- Custom designed ROBIN card with 3 * 64 MB RAM
- 3 S-Link fiber optic ports
- IPMI 2.0 BMC
- Redundant (2 out of 3) power supply
- SuperMicro X6DHE-XB 333 MHz FSB 512 MB RAM
- PCIe based 4-port NIC from Silicom
- Single Intel Xeon @ 3.4 GHz (Irwindale)
- SuperMicro X6DHE-XB 333 MHz FSB 512 MB RAM
- PCIe based 4-port NIC from Silicom
- Single Intel Xeon @ 3.4 GHz (Irwindale)
- SuperMicro X6DHE-XB 333 MHz FSB 512 MB RAM
- PCIe based 4-port NIC from Silicom
- Single Intel Xeon @ 3.4 GHz (Irwindale)
- SuperMicro X6DHE-XB 333 MHz FSB 512 MB RAM
- PCIe based 4-port NIC from Silicom
- Single Intel Xeon @ 3.4 GHz (Irwindale)
- SuperMicro X6DHE-XB 333 MHz FSB 512 MB RAM
- PCIe based 4-port NIC from Silicom
Hardware Components

68 ROS PC (liquid Argon sub detector)
Operational Monitoring

- Lowest level: IPMI & ssh
- Server level: Nagios
- User level: Web browser

Features:
- History charts
- Automatic E-mail notification in case of problems

System status

Service Status Details For Host 'pc-tll-ros-bc-01'

Features:
- History charts
- Automatic E-mail notification in case of problems

System temperature

Fan speed
## Hardware reliability

<table>
<thead>
<tr>
<th>Type of component</th>
<th>Number of installed units</th>
<th>Number of broken units</th>
<th>Failures per year [%]</th>
<th>Failures in 2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC Motherboard</td>
<td>150</td>
<td>3</td>
<td>0.77</td>
<td>1</td>
</tr>
<tr>
<td>CPU</td>
<td>150</td>
<td>1</td>
<td>0.26</td>
<td>0</td>
</tr>
<tr>
<td>Memory DIMM</td>
<td>300</td>
<td>3</td>
<td>0.39</td>
<td>1</td>
</tr>
<tr>
<td>Power Supply module</td>
<td>450</td>
<td>4</td>
<td>0.34</td>
<td>1</td>
</tr>
<tr>
<td>IPMI BMC</td>
<td>150</td>
<td>4</td>
<td>1.03</td>
<td>0</td>
</tr>
<tr>
<td>CPU ventilator</td>
<td>150</td>
<td>2</td>
<td>0.51</td>
<td>0</td>
</tr>
<tr>
<td>chassis ventilator</td>
<td>450</td>
<td>1</td>
<td>0.09</td>
<td>0</td>
</tr>
<tr>
<td>4-port NIC</td>
<td>150</td>
<td>1</td>
<td>0.26</td>
<td>1</td>
</tr>
<tr>
<td>ROBIN cards - broken</td>
<td>614</td>
<td>23</td>
<td>1.45</td>
<td>1</td>
</tr>
<tr>
<td>ROBIN cards – intermittent errors (Firmware issues)</td>
<td>614</td>
<td>36</td>
<td>2.26</td>
<td>15</td>
</tr>
</tbody>
</table>

Average age of the hardware: 2.6 years
System integration issues

- The individual PSUs of the PCs generate a significant inrush current peak when power is restored after a power cut.
- This may cause breakers in the rack to trip.
- First solution: Power staggering barrettes
  - But recent questions on long-term reliability
- So now plan second solution
  - Thermistors
The application retrieves data fragments from the ROBINs, combines them in a unique fragment and sends it to L2/EB.

Multi-threaded C++ program running under Linux (SLC4)

ROS Application

- Request handlers
- Data requests from HLT
- Clear requests from DFM
- Request receiver
- ROS request queue

Data to HLT

PowerPC processor in ROBIN runs C program booted from FLASH memory

ROBIN

- Event store
- PPC
- FPGA

ROBIN request queues

PCI bus

TIPP09 Markus Joos – The ATLAS Read-Out System: performance with first data and perspective for future
System performance in 2008

- The ROS PCs were powered almost permanently.
- Most of the time they were used for data taking with a trigger on cosmic events.
- ROS data output statistics for August to October:
  - ~900 TB of data
- All detectors were commissioned at full S-Link (ROD-ROS) speed.
- During selected periods high-rate tests were performed with pre-loaded data (ROBIN not involved).
  - 136 ROS PCs delivered 5.3 GB/s to HLT (40 MB/s/ROS).
Performance of the standard ROS

Test Setup

- NIC
- PCI
- ROBIN
- ROBIN
- ROBIN
- ROBIN

GbEth.

Requester PCs, emulate HLT nodes

Lab measurement
Differences wrt deployed system:
- All network traffic based on TCP/IP
- ROBINs generate data internally

- Canonical fragment size in ATLAS: ~ 1 kByte (256 words)
- Meets the original requirements (~20 kHz L2 rate) for small fragment sizes
- Too slow for large fragments
- Bottleneck seems to be the ROS CPU (ROBINs & NICs can sustain much higher rates)

![Graph showing performance metrics]

- 3 KHz EB
- 3 ROLs per L2 request
- 2 GBE links
- Default system
- 100 kHz L1 rate
- TCP/IP only

- External L2 request rate (kHz)
- Event fragment size (32-bit words)
- Total transfer bandwidth (MB/s)

- L2 rate
- Bandwidth
- S-Link limit

TIPP09  Markus Joos – The ATLAS Read-Out System: performance with first data and perspective for future
The original configuration of the OS & drivers turned out to be inefficient. Finally we obtained the best performance by:

- Turning hyperthreading off
- Using a uni-processor kernel
- Tuning the interrupt coalescence of the network driver
- Changing the SELinux configuration

- Performance is now OK for all fragment sizes
- However would like more headroom
Higher ROS performance - motivation

- ATLAS (upgrade) phases
  - Phase 0 (until 2013, luminosity: up to $1 \times 10^{34}$ cm$^{-2}$ s$^{-1}$)
    - Need more ROS performance to:
      - have headroom for ROS PCs with high L2 request rates
      - compensate for higher rates due to modified thresholds of the L2 trigger
      - allow for additional bandwidth-demanding types of triggers. E.g.:
        » Inner detector full scan for b-physics
        » Calorimeter full scan for missing $E_T$
  - Phase 1 (2013 – 2017, luminosity: up to $3 \times 10^{34}$ cm$^{-2}$ s$^{-1}$)
    - Higher data rates due to increased luminosity
    - Still use (current) ROS PCs & ROBINs
    - Requires more network bandwidth (switches, ROBINs & ROS)
  - Phase 2 (from 2018, luminosity: up to $10 \times 10^{34}$ cm$^{-2}$ s$^{-1}$)
    - Much higher data rates
    - Replace ROS system
Higher ROS performance - options
(for phase 0 & 1)

- The main bottleneck of the ROS is the network interface to the HLT
  - Only 2 GBE links per ROS
  - Network protocol (mix of UDP and TCP/IP) handled by the CPU of the ROS PC
- 3 approaches to solve the network limitation

Install smart NICs (to offload CPU from the TCP/IP protocol)

Replace the motherboard, CPU and memory of the ROS PCs with faster hardware and connect additional GBE lines from the ROS PCs to the HLT network

Connect the ROBINs directly to the HLT network

Chelsio S320E 2-port smart NIC

SuperMicro X7DB8-X MB with 2 * 2.66 GHz quad core Xeon and RAM @ 667 MHz

This may be one of the last MBs with >3 64bit PCI slots -> Development of PCIe based ROBIN started

GBE port of ROBIN (UDP only)
Impact of the Smart NIC

Laboratory measurement
- TCP/IP only

- No significant gain in performance
- Driver for smart NIC not yet tuned for best performance (waiting for optimized driver from manufacturer)
- More work required but overall potential seems to be low
Impact of faster MB, CPU & RAM

Test Setup

Requester PCs, emulate HLT nodes

fast ROS PC

GbEth.

NIC

PCI

ROBIN

ROBIN

ROBIN

ROBIN

NIC

Test Setup

Requester PCs, emulate HLT nodes

- L2 request rate (almost) fragment size independent
- L2 request rate increases by 50% to 150%
- (expensive) ROBIN cards can be reused

Best performance with 2 (of 8) cores
-> effect of (extensive) use of mutexes?
Impact of Read-Out via PC & ROBIN cards

Full system test with (small HLT) farm
Only preliminary results so far

- Functionality OK
- current test system limits performance

- This ROS configuration has the potential to deliver more performance than the ROS with the faster motherboard & CPU
- Further optimization of the system (software) required
Summary and Conclusions

- Since its installation in 2006/2007 the ROS system has worked very reliably
- The ROS in its current configuration meets the requirements that were specified in the ATLAS Technical Design Report
- Several alternatives exist for the further improvement of the performance
  - More detailed tests have to be carried out in the deployed system to better understand the relative advantages and disadvantages of these alternatives
- The development of a PCIe based ROBIN has been started
  - Because motherboards with at least 4 64-bit PCI slots become difficult to find
  - Faster PPC CPU will also improve ROBIN performance
- Based on today’s understanding of the ATLAS TDAQ (HLT rejection factor and algorithms) as well as the planned upgrades of ATLAS and LHC the current ROS architecture fulfills the requirements of phase 0 & 1 of ATLAS