The Use of Commodity Products in the ATLAS Level-2 Trigger

M. Dobson (CERN) for the ATLAS level-2 trigger groups

February 8, 2000
CHEP 2000, 7–11th February, Padova, Italy

- Introduction, Architecture and Required Performance
- Testbeds
- Component Performance
- System Scaling
- Conclusions

These are partial measurements and the results have to be finalised.

Introduction

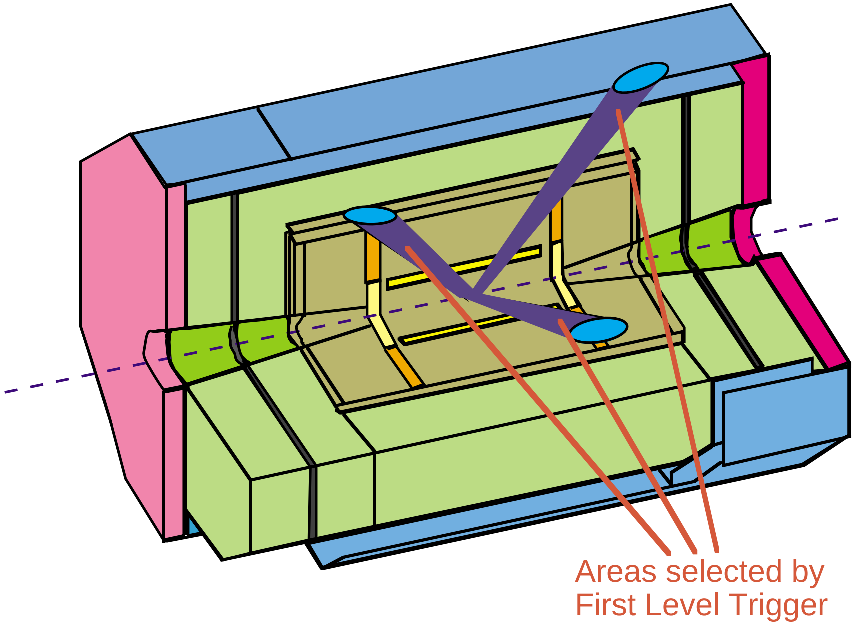
ATLAS three-tier Trigger:

- Level-1: custom hardware, reduce rate 40 MHz to < 100 kHz.
- Level-2 and Event Filter (EF): reduce rate to about 100 Hz.

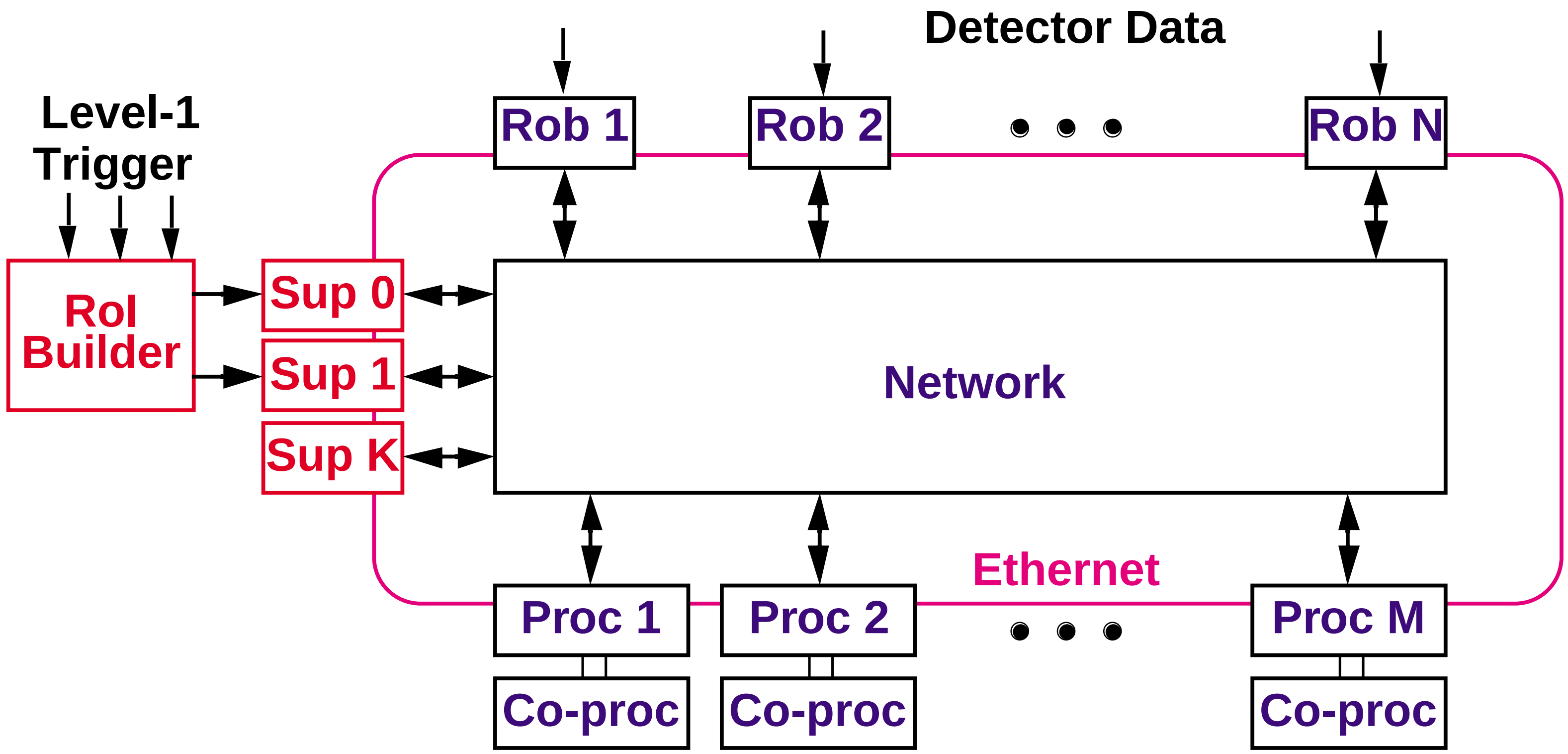
The Level-2 Trigger:

- Event data (1–2 Mbyte) for level-1 accept stored in ReadOut Buffers (ROBs).
- Region of Interest (RoI) guidance from level-1 reduces network bandwidth and processing power required (few percent of event data only).
- Sequential Selection Strategy: rejection at different stages in processing.
- Commodity Components wherever possible: PCs, OS, NICs, Switches.

Regions of Interest (RoI)



The Level-2 Trigger Architecture



Characteristics:

- RoI Builder: custom hardware; combine RoI pointers into event record.
- Supervisor farm: commodity processors; receive event records from RoI Builder; allocate each to a processor; receives trigger decision; distributes reject decisions to ROBs.
- Processor farm: commodity processors; receive event record; in one or more steps request data from ROBs, process it and make decision; send final trigger decision to supervisor farm. Possible use of co-processors (FPGAs).
- ROBs: respond to data requests; clear events as indicated by supervisor.
- Network: transport data and messages between components.

The Required Performance

Requirements from so-called Paper models (http://www.nikhef.nl/pub/experiments/atlas/daq/modelling.html).

Paper models: simple spreadsheet models, calculating average rate, bandwidth, latency and load for components in the system. Does not account for queueing effects.

Supervisor Farm	Processor	ROB	Network
Operation up to 100 kHz. Input from RoI Builder and level-2 processors	Few hundred processors Event rate < 1 kHz per processor Bandwidth of order 10 MByte/s per proc.	$\simeq 1600$ ROBs Data request rate: of order 1–10 kHz Level-2 bandwidth out: 1–10 MByte/s per ROB	Aggregate bandwidth: Data: $\simeq 5.3$ GByte/s Control messages: $\simeq 250$ MByte/s Ports: $\leq$ 2350

Processor (data collection) and ROB performance dependent on I/O overhead $\Rightarrow$ low latency communication drivers needed.

The Testbeds

Aims:

- measure component performance and check against requirements
- study scaling and performance with system size
- provide data for full system computer models

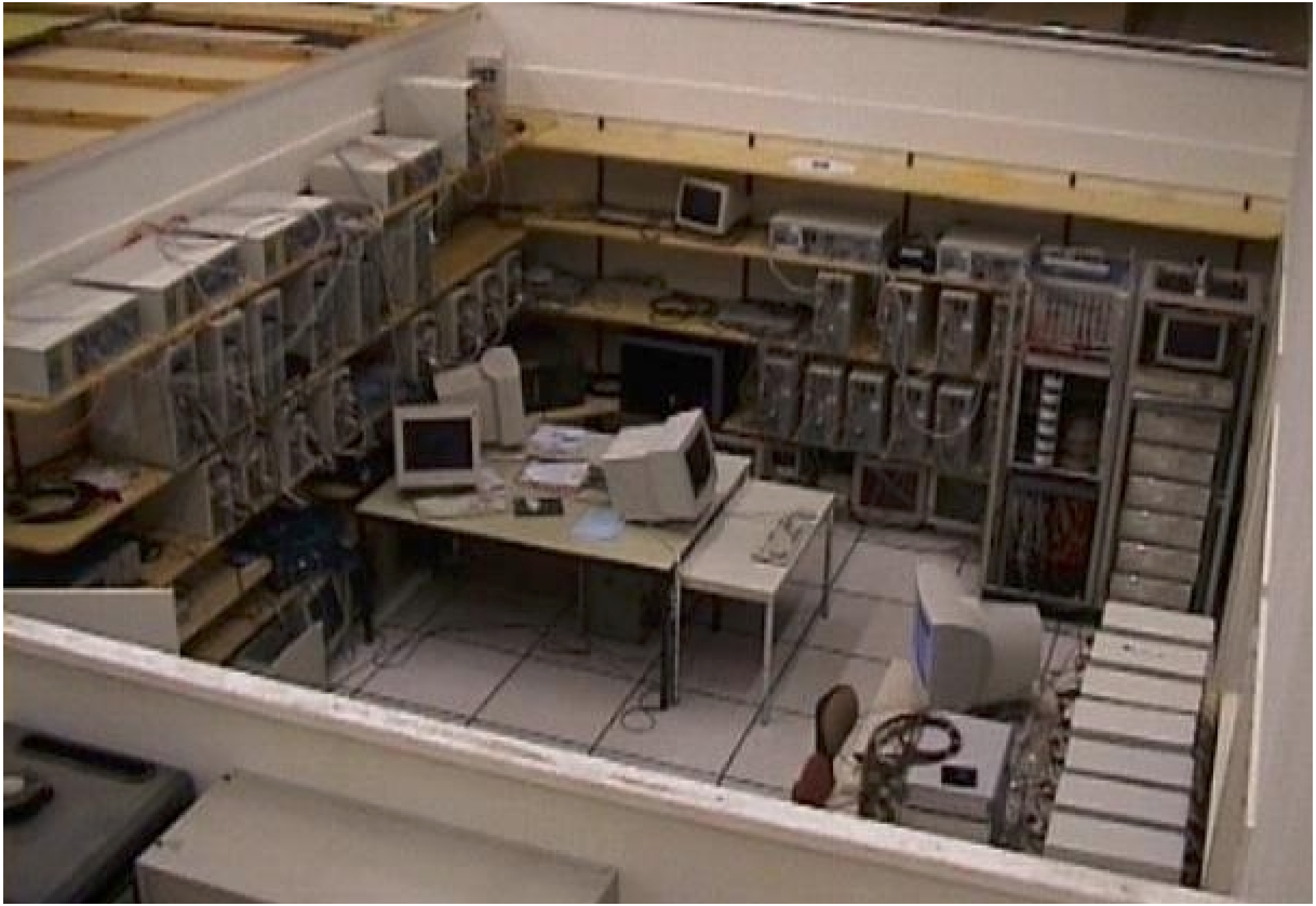
Characteristics:

- 25–50 nodes (few percent of system).
- Ethernet, ATM, and SCI networks.
- Hardware shared between testbeds for different technologies.
- ATM 48-port, 155Mbit/s FORE switch.
- Ethernet: 3 BATM Titan T4 Fast/Gigabit switches (32 Fast or 4 Gigabit ports per switch).
- SCI: 16 port Dolphin switch ($\simeq$ 6.0 Gbit/s switch).
- SCI on commercial cluster: 96 nodes at Paderborn University.

Software:

- OO C++ prototype level-2 software (not yet optimised) called reference software.
- OS: Linux, Windows NT, Solaris.

The Testbeds



The Supervisor and RoI Builder

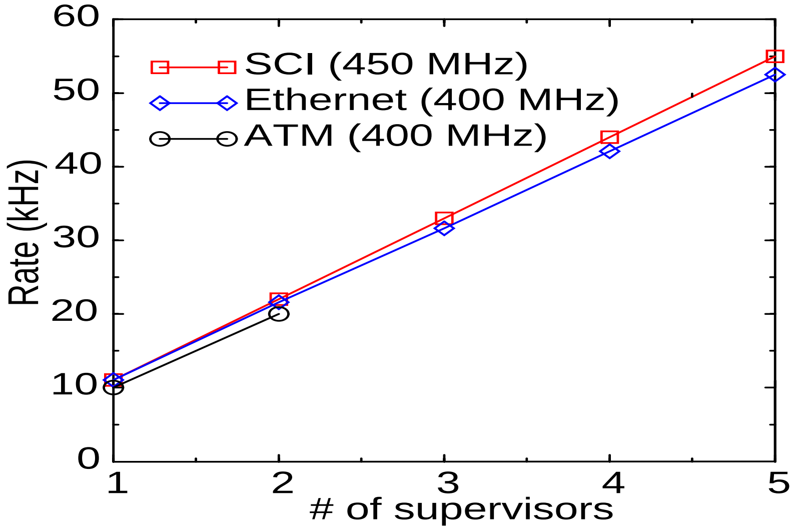
Supervisor tests use a Supervisor emulator running on a PC, with no RoI builder connection.

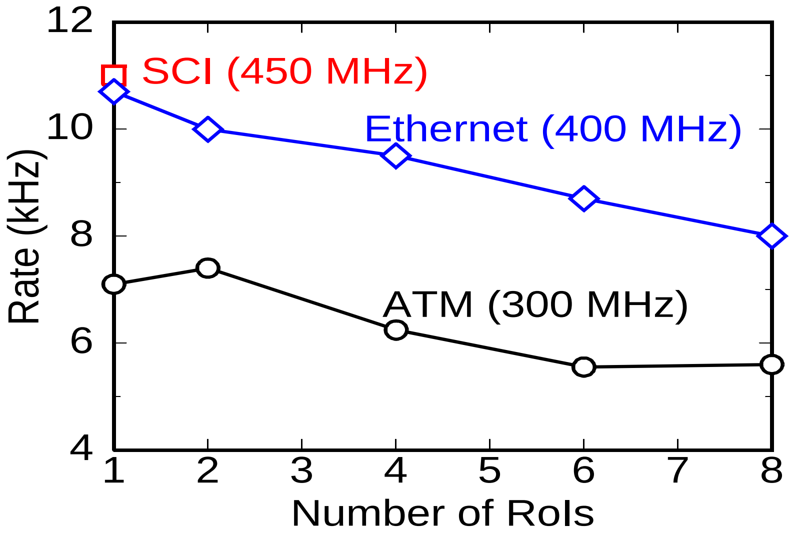
Supervisor Emulator Performance:

- 10 kHz per supervisor processor for 1 RoI per event
- use of multicast: reduce traffic, rate independent of # ROBs
- rate versus # level-2 data processors linear until saturation
- linear scaling of rate versus # supervisor processors

Demonstrated that a small farm of order 10 processors is sufficient for the supervisor task.

Still better performance shown with RoI builder and earlier optimised software for ATM testbed.



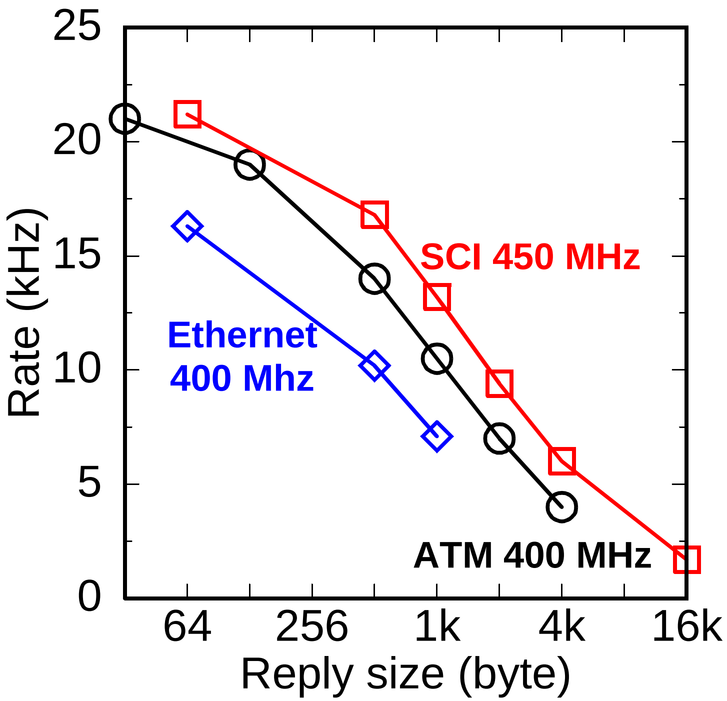


The ReadOut Buffer

ROB access performance tests use ROB emulator on PC.

For a typical data size of 1 kByte, a maximum service request rate of 12 kHz is needed.

The emulator has achieved up to 13 kHz with the current technology and non-optimised software.



ROB prototype hardware has demonstrated 30 MByte/s and 50 kHz maximum fragment request rate in separate tests.

⇒ ROB service rate/bandwidth acceptable for up to 100 kHz level-1 rate.

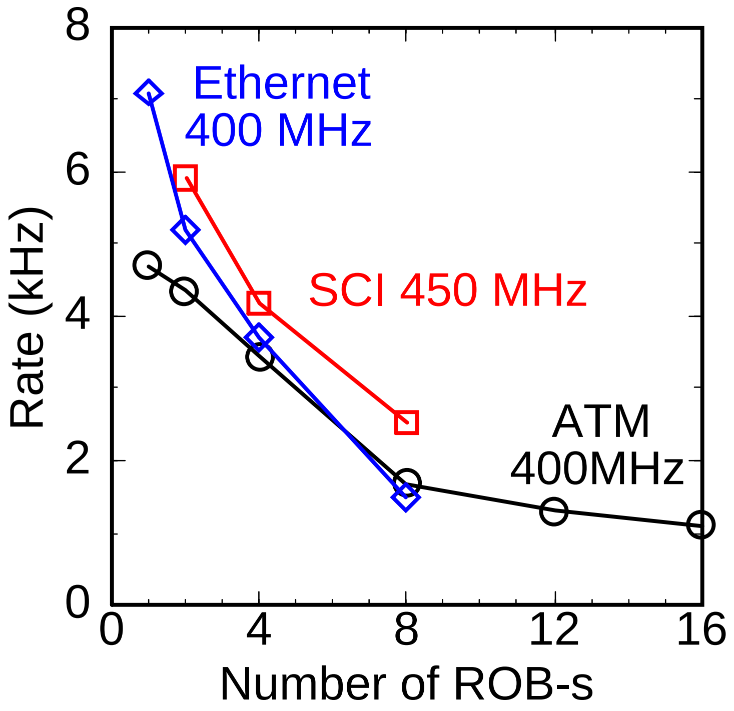
The Processor

Emphasis, in studies described here, is placed on data collection task, and in particular on RoI data collection.

RoI data collection is from many ROBs (3 to 20 depending on subdetector).

In all cases, RoI data collection does not occupy large fraction of the processor time.
Rate required by paper models is < 1 kHz.

Complete subdetector scan (collection from several hundred ROBs) also needed by some algorithms. No tests performed here.



Performance conclusions:

- meets requirements for rate and bandwidth (100 kHz level-1 rate) for data collection with a few hundred processors.
- we believe it will be able to meet requirements for full level-1 rate with algorithms when system comes on line.

FPGA co-processors for associated computing intensive algorithms, if necessary.

The Network

Three technologies have demonstrated the required performance (for the small systems used).

Ethernet and ATM drivers are custom written, SCI has used optimised commercial drivers.

Tests have demonstrated the use of modest size switches up to about 1 GByte/s.

There are many possibilities to construct large networks with commodity technologies; switches which can handle many 10's of Gbit/s are currently available.

Conclusions:

Market offers large choice of performance items. Modelling efforts are going on to study large networks built from small size switches.

System Performance and Scaling

Large system operation on the Paderborn cluster (96 nodes):

- approximately linear scaling of system rate with # of ROB/processor pairs
- stable performance of supervisor versus # processors
- correct operation of software on significant size system

Sequential Selection:

- reduce network bandwidth and processor requirements
- allow more complex algorithms to be run at a low rate, while retaining a short average latency.

Tests have been run to demonstrate the feasibility of such selection strategies.

Conclusions

Level-2 strategy and architecture implemented on moderately large testbeds.

Performance of components consistent with or close to requirements from the paper models.

Many components use commodity products (OS, processors and network technology): performance is adequate.

Custom items may be needed for the drivers and co-processors in addition to the reference software.

RoI Builder implemented in custom hardware, because of need to combine high rate data streams. Performance of prototype components already meet maximum level-1 rate.

Too early to conclude on role of commodity items for ROBs.

The performance obtained in these tests is with non optimised software. Optimisations will give still better results.

Final results can be found in the forthcoming document: “ATLAS DAQ, High-Level Triggers and DCS Technical Proposal”.

Further Work

Optimise Reference Software.

Evaluate system with prototype algorithms and simulated data.

Better integration of ROB hardware into system.

Examination of benefits of SMP machines for ROB and Processor.

Integrate with other prototype Trigger and DAQ components.

Large network evaluation and modelling, plus cost estimate for different technologies.

Better use via drivers and hardware of fast link speeds.

Operation on larger cluster.

Still need to show system scaling on a larger system.