

Using SCI to Implement the Local-Global Architecture for ATLAS Level 2 Trigger

Involving:

A. Bogaerts, E. Denes, F. Giacomini, R. Hauser, P. Werner
CERN, CH-1211 Geneva 23, Switzerland

R.E Hughes-Jones, S.D. Kolya, D. Mercer
The University of Manchester, Manchester, M13 9PL, UK

D. Botterill, R.P. Middleton, F.J. Wickens
Rutherford Appleton Laboratory, Chilton, Oxon, OX11 0QX, UK

M. Liebhart
TU-Graz, Institute for Technical Informatics, 8010 Graz, Austria

J. Dawson, J. Schlereth
Argonne National Lab, Illinois, USA,

A. Guglielmi
DEC Joint Project Office, CERN, 1211 Geneva 23, Switzerland

R. Hughes-Jones

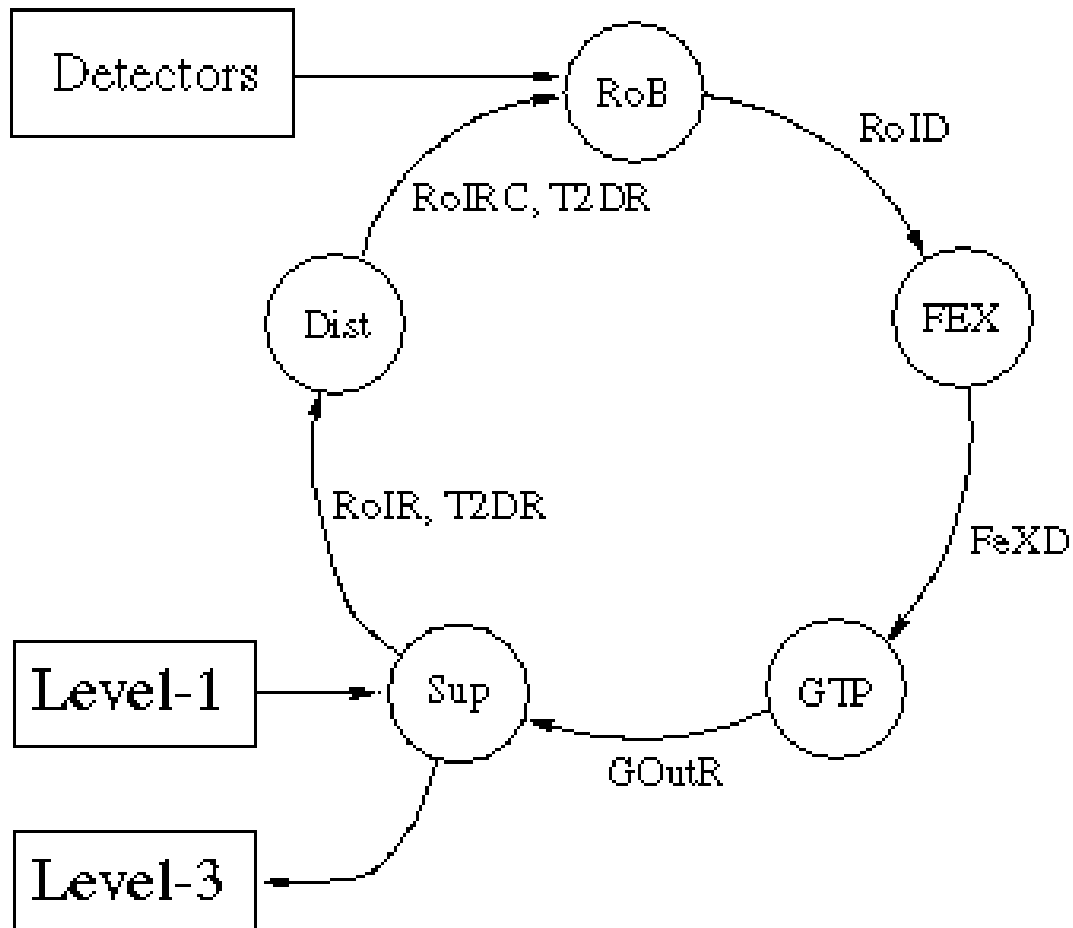
February 1998 CERN

Aims of the Project:

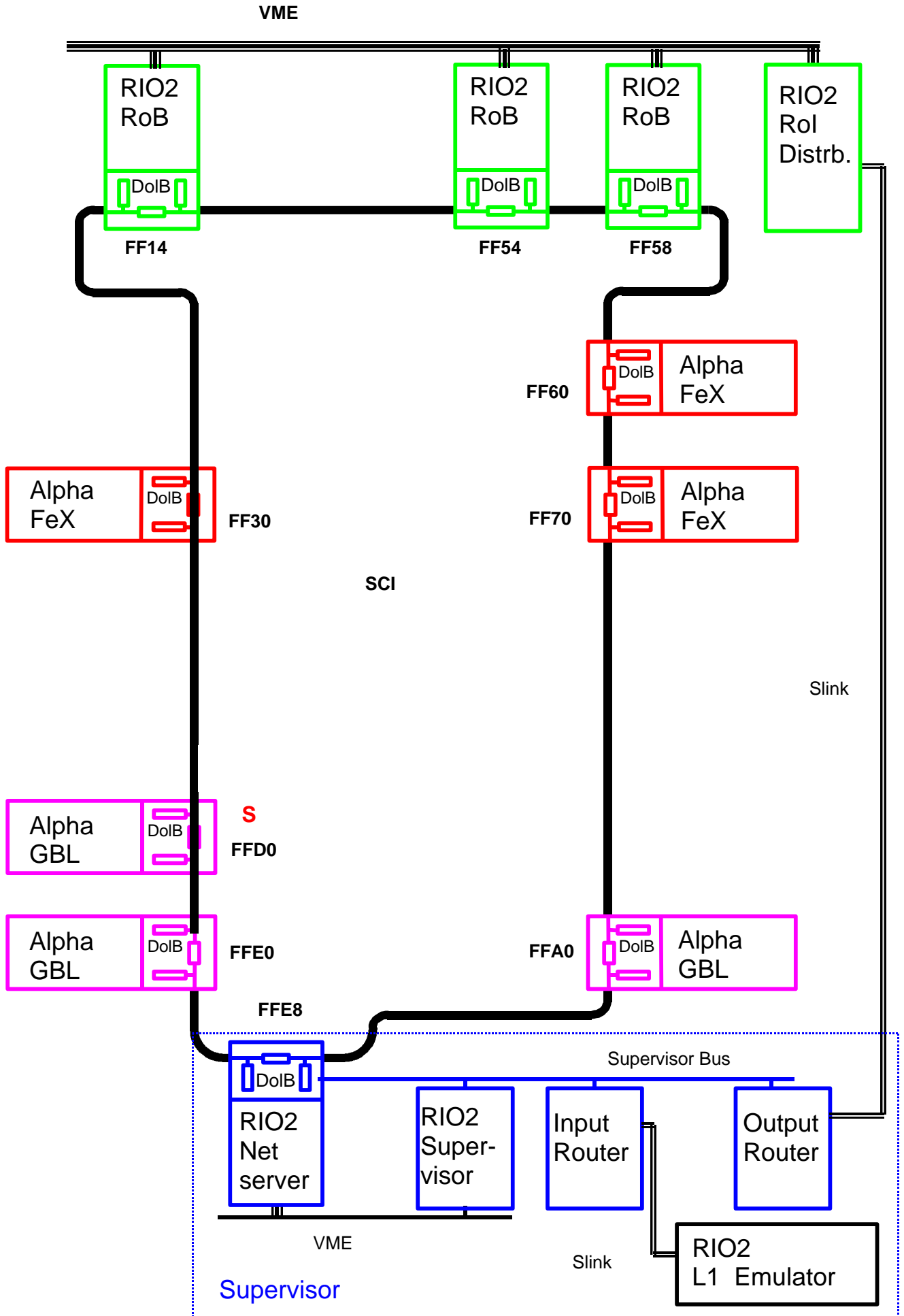
- To implement and demonstrate the operation of a vertical section of the Level 2 trigger Local-Global architecture using SCI interconnect.
- To show the general suitability of SCI for the ATLAS Level 2 trigger by including:
 - SCI using DMA-Ring buffer mode
 - SCI using Transparent mode
 - Include the 4-port SCI switch
- To demonstrate that the message protocol is sufficient and stable.
- To measure the performance of the components and the system implementation.
- To include error detection and rudimentary recovery.
- To investigate the interface between the high performance trigger functions and the essential DAQ functions such as initialisation, control, and monitoring.
- To implement the system simply and efficiently, but with flexibility and portability.

The Key Points of the Architecture

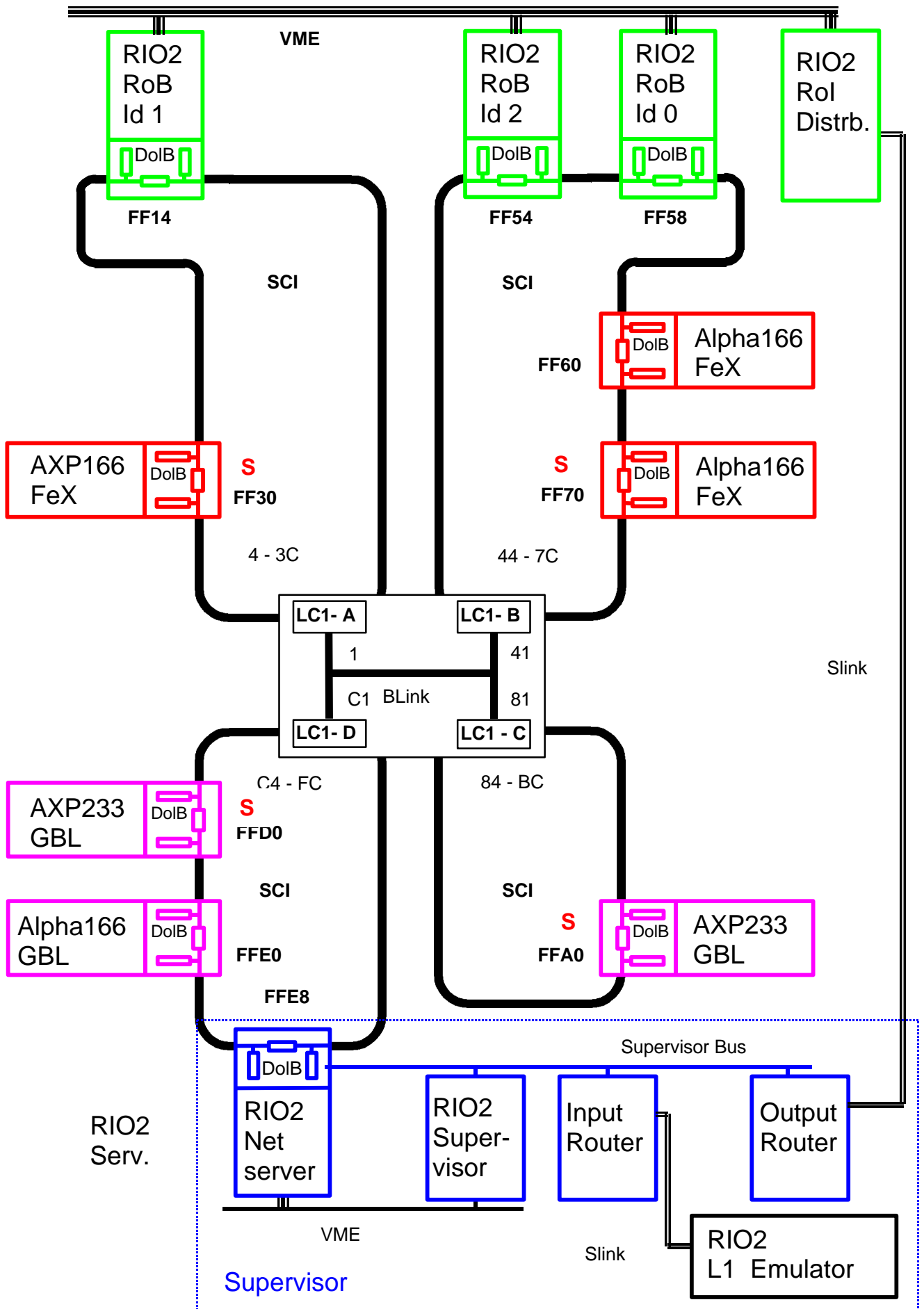
- Requested push, the buffers send data on a request from the supervisor.
- The supervisor controls and accounts for all events.
- The data messages that are passed round the system contain their own destination address or routing information.
- All the data from 1 RoI for 1 sub-detector is sent to 1 FeX.
- This is the local processing.
- Many FeXs are working in parallel on 1 event, and send their results to 1 Global processor.
- Many events will be in the system at one time.



SCI Vertical Slice 3 RoB, 3 FeX, 3 Global, Supervisor



SCI Vertical Slice 3 RoB, 3 FeX, 3 Global, Supervisor 4*4 Switch



Messages in the Loop

| | |
|---|---------------------------|
| Supervisor processor locates the L1 Rol data Sends to Output Router [RoIR] | Super bus |
| Output Router sends Rol information to Rol Distributor [RoIR] | Slink |
| Rol Dist. Sends Rol information to RoB [RoIRC] | VME |
| RoB locates the detector data, sends RoID to FeX [RoID] | SCI |
| FeX builds trigger event from n RoBs Sends FeXD to GBL processor [FeXD] | SCI |
| GBL builds trigger event from n FeXs Sends decision to Network Server [GoutR] | SCI |
| Network server sends decision to Supervisor | VME |
| Supervisor Checks event id and type in table and prepares decision for RoBs [T2DR] | Super bus Slink VME |

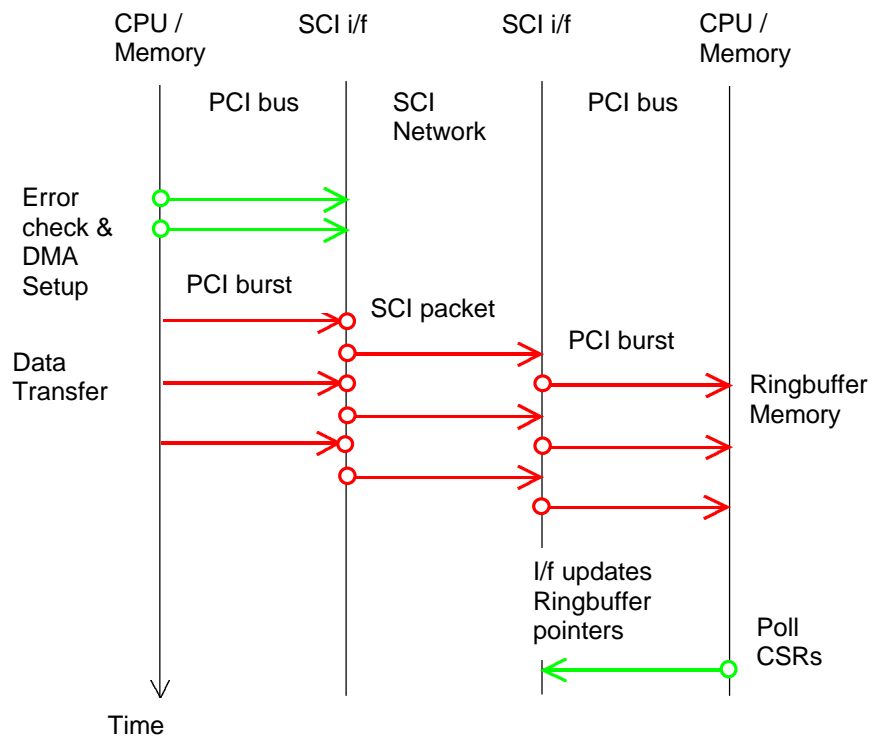
Hardware

- | | | |
|-------------------------------|---------|--------|
| – Supervisor + Network server | PowerPC | LynxOS |
| – RoI Distributor | PowerPC | LynxOS |
| – Up to 3 RIO-RoB Emulators | PowerPC | LynxOS |
| – Up to 3 FeXs | Alpha | µC/OS |
| – Up to 3 Globals | Alpha | µC/OS |
| – 4-port SCI Switch | | |

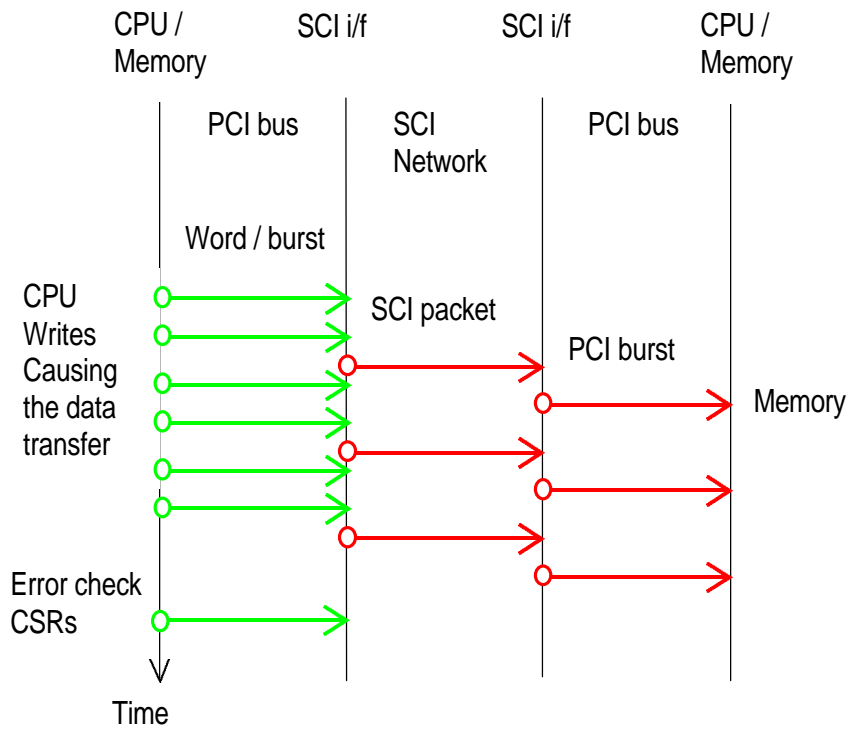
Software Infrastructure

- | | |
|-------------------|--|
| – RT Kernel | - µC/OS & LynxOS (threads, queues, mailboxes, etc...) |
| – Control network | - raw ethernet protocol |
| – Error Reporting | - EMU |
| – SCI libraries | - Simple API for DMA-Ringbuffer - Based on IEEE 1596.9 (draft) for Transparent |
| – Data Transport | - DMA simple use of API - Transparent Message Passing layered on SCI library |

Use of CPU, PCI, and SCI for DMA-Ringbuffer Mode

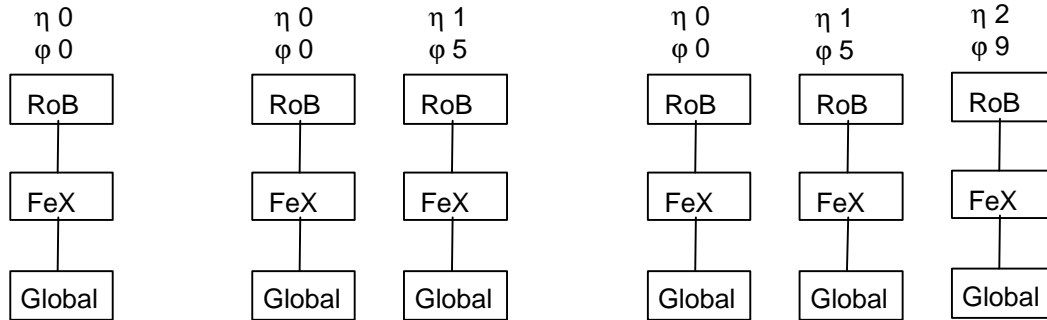


Use of CPU, PCI, and SCI for Transparent Mode

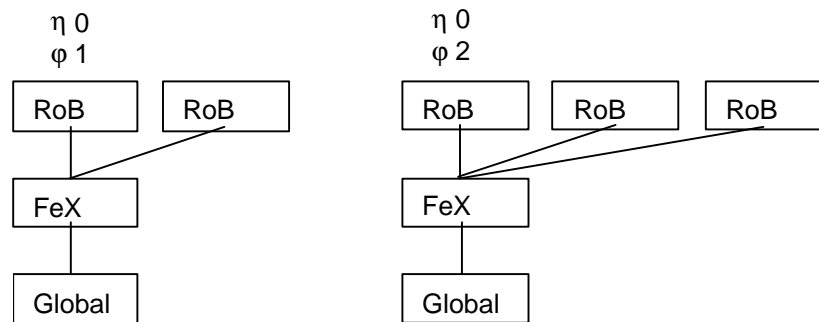


Topologies Tested

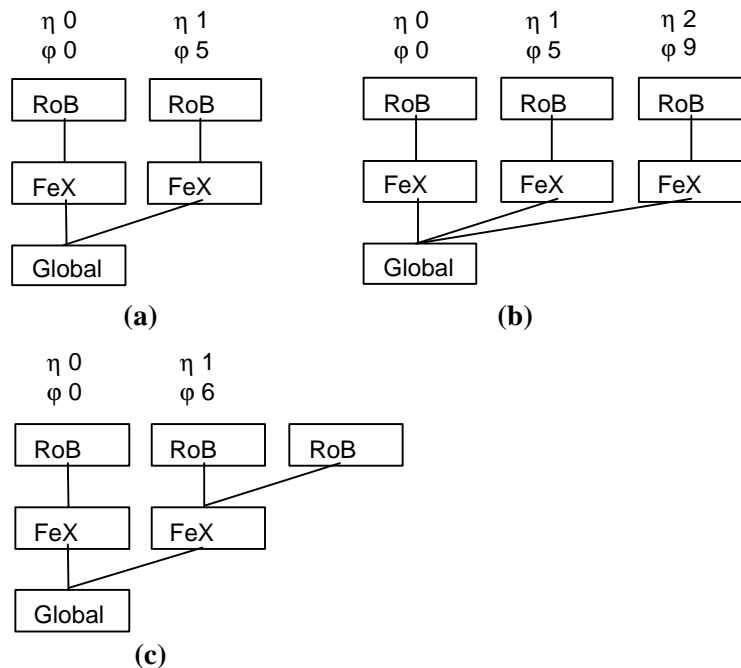
- A single stripe1 RoB 1 FeX 1 Global.
- Parallel stripes with 2 or 3 RoB-FeX-Global



- Trigger Fragment building in the FeX, multiple RoB per RoI.



- RoI Parallelism, multiple RoB-FeX channels feeding 1 Global processor.



The Measurements Made

The measurements were made with the loop closed and under the control of the Supervisor operating in "free running" mode.

The Supervisor initiated a new event as fast as it could, provided the Max No. of events in the loop was not exceeded.

For each configuration measured:

1. Loop latency -Time taken for an event to be transmitted though the system. Measured as follows:
Supervisor : start time - Form RoIR
stop time - T2 decision validated
2. Average time between event decisions:
Total number of events / elapsed time
3. Times taken by the internal parts of Supervisor, RoB complex, FeX, & Global

These were measured as:

- Function of the size of the data transferred Rob -> FeX
- Function of the maximum no. events the Supervisor will allow in the system.

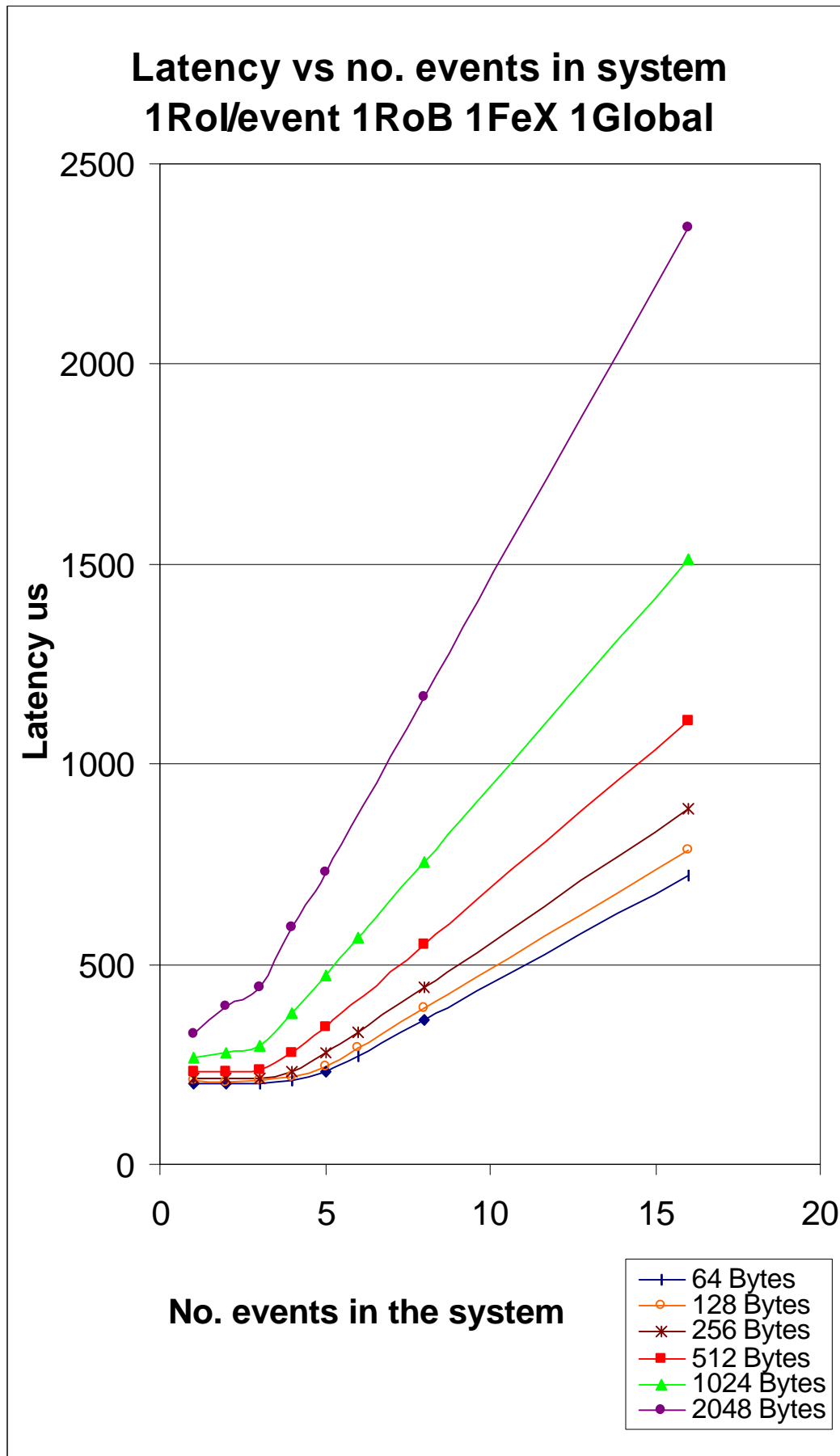
Each configuration used $\sim 10^6$ events or ~ 1 Billion events in total.

A logic analyser was used to observe the traffic on the FeX and Global PCI bus and at various points of the SCI ring.

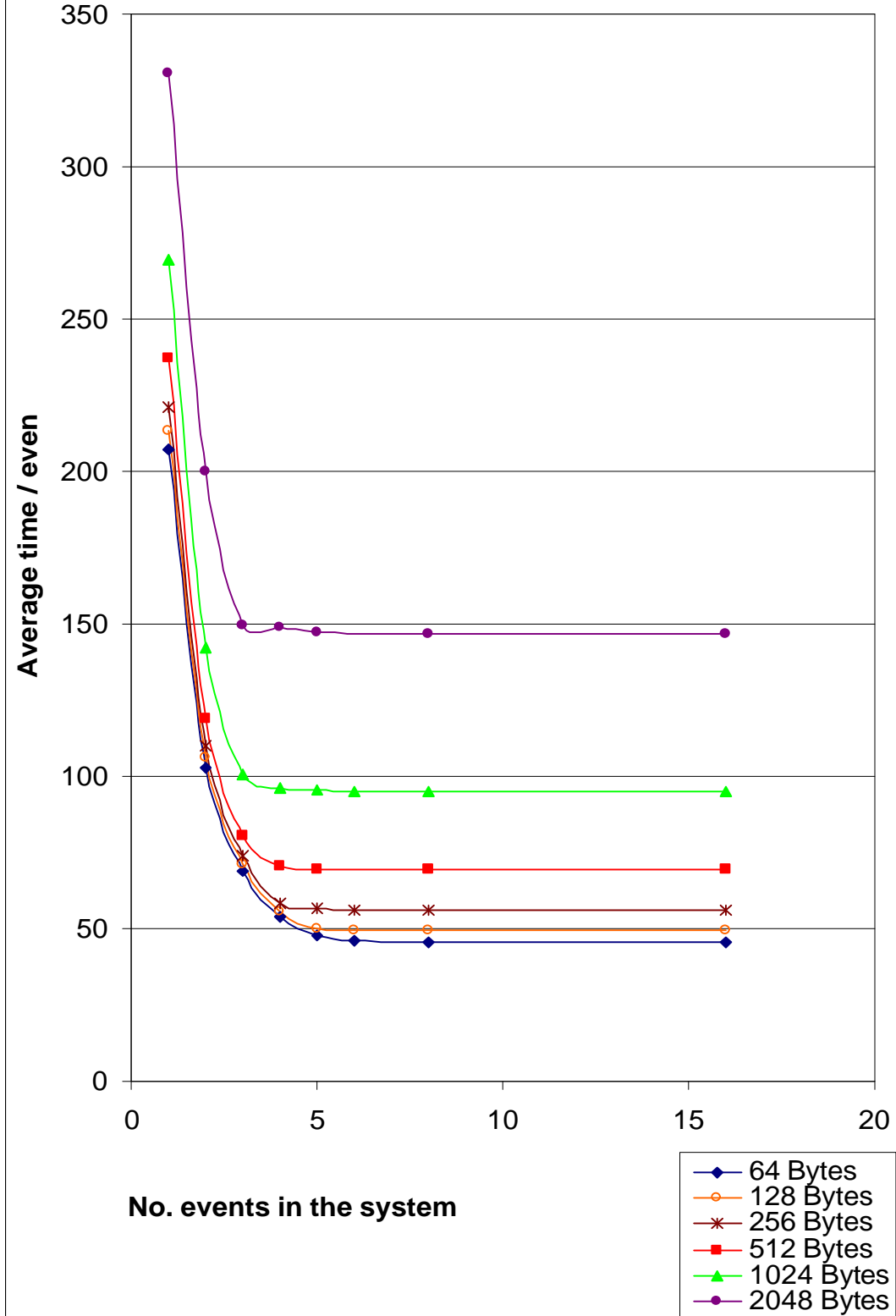
When events queue in the system:

- Ave Time between events = Time of slowest element
- Latency = No of events in loop * Time of slowest element

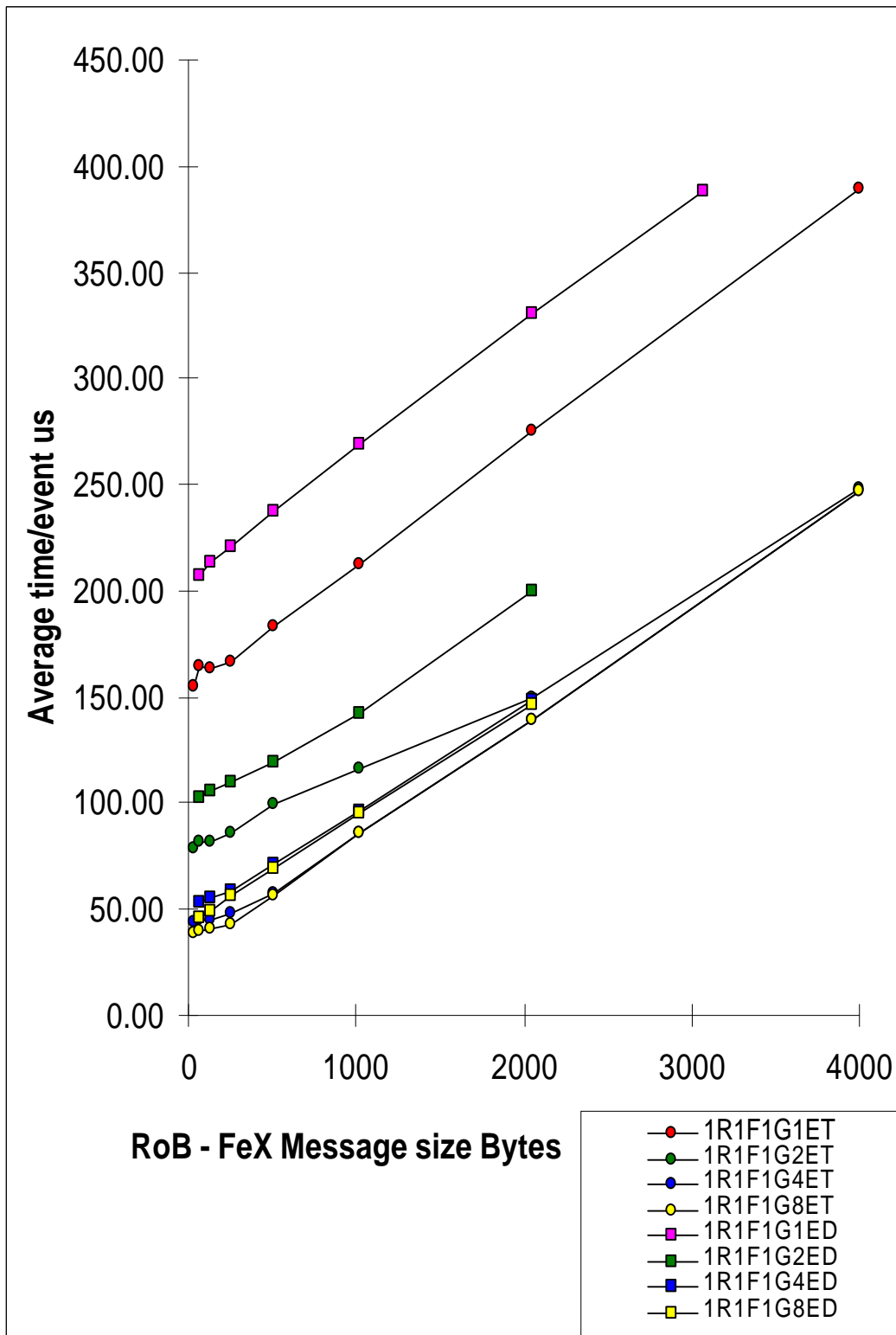
Measurements with the Vertical Slice



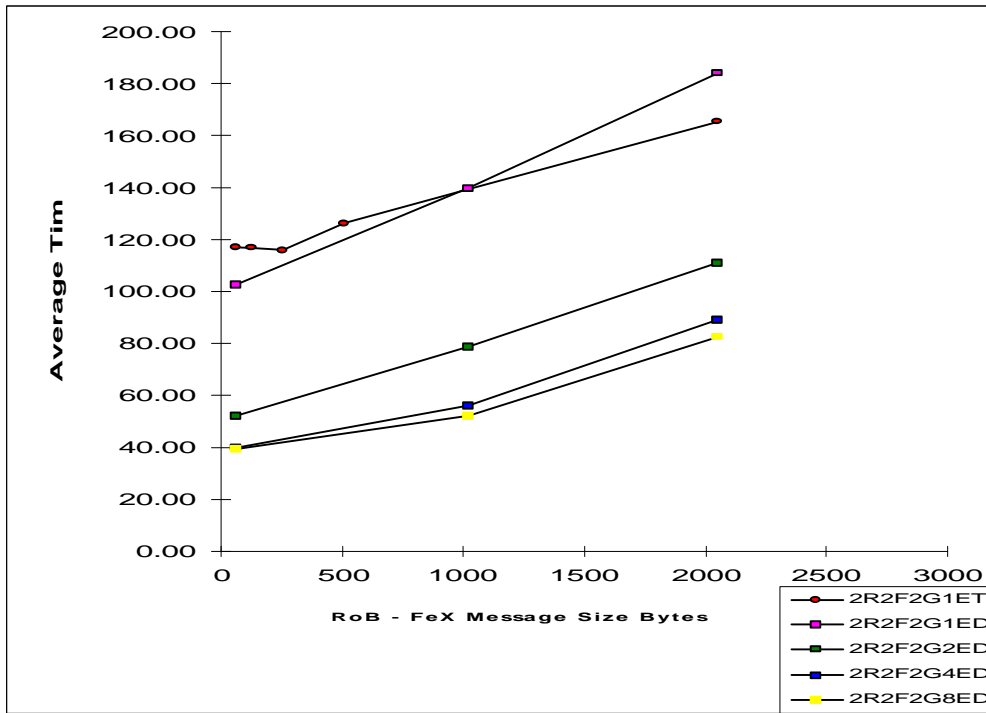
Ave. time / event vs no. events 1RoI/event 1RoB 1FeX 1Global



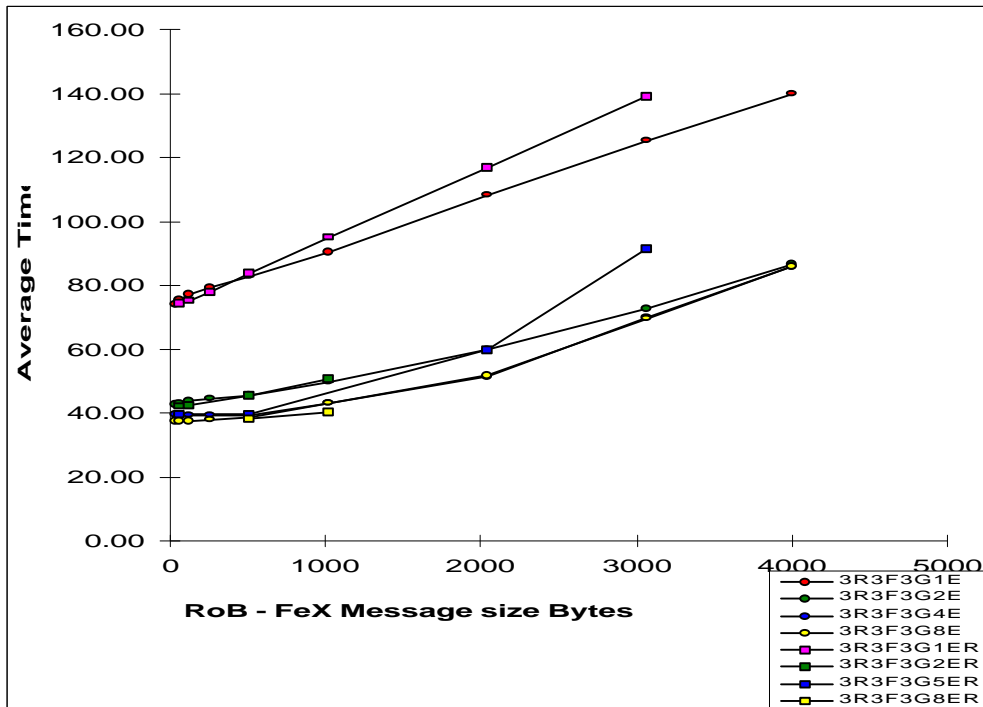
The Average time per event as a function of the RoB – FeX messages size.



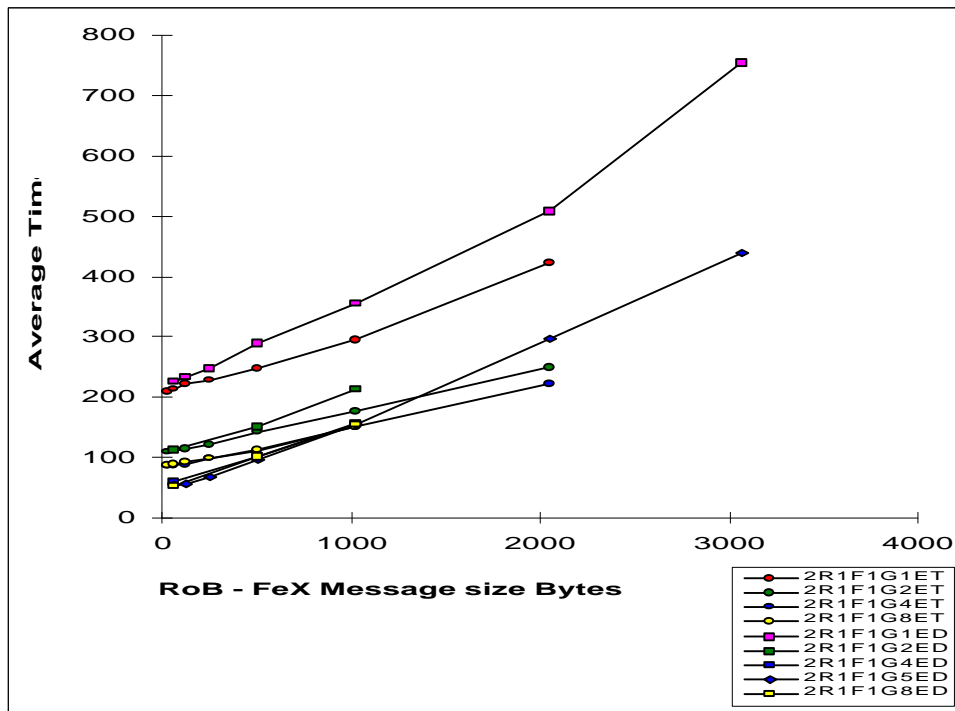
Two Parallel RoB FeX Global Stripes



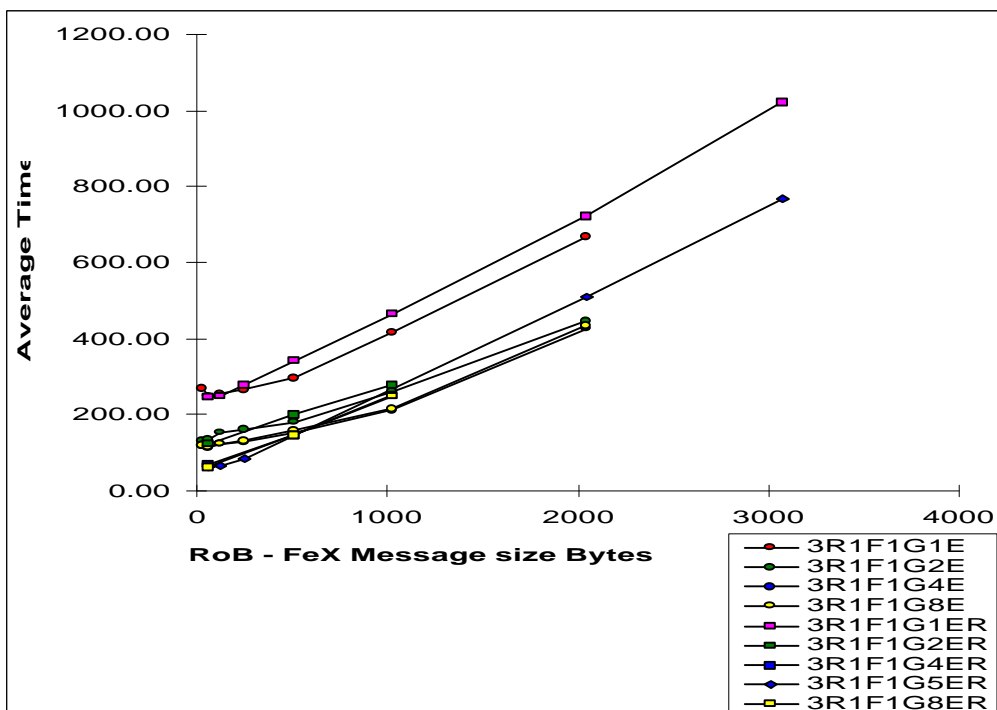
Three Parallel RoB FeX Global Stripes



Trigger Fragment Building 2RoBs 1FeX 1Global



Trigger Fragment Building 3RoBs 1FeX 1Global



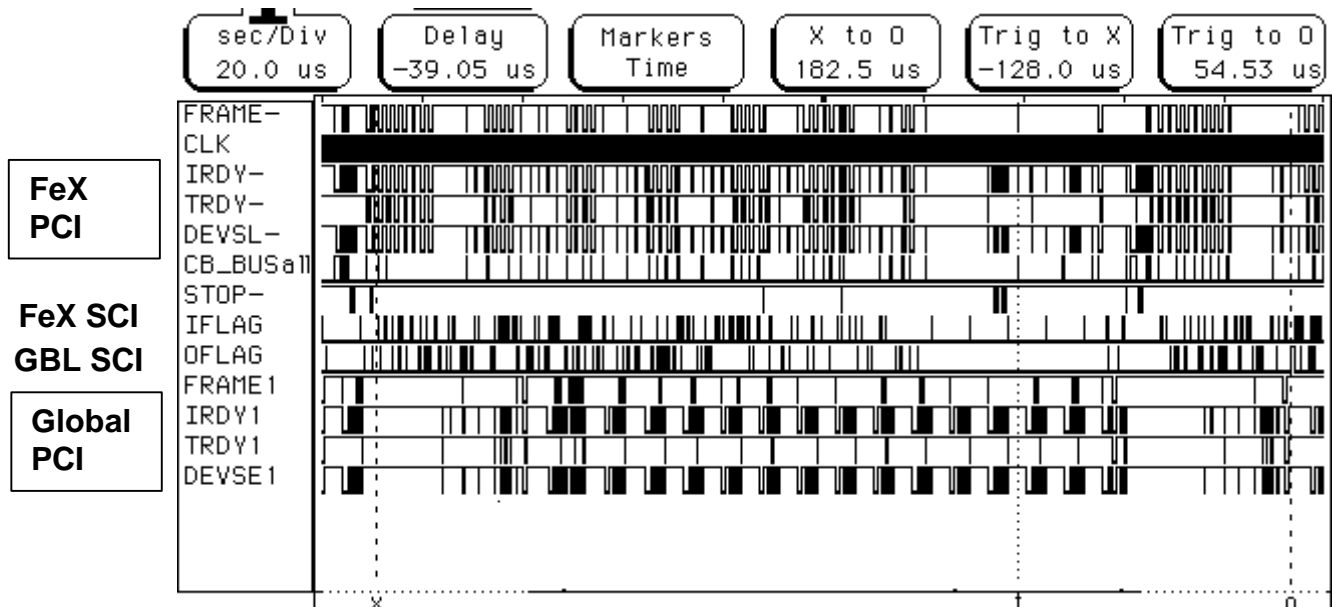
Some Key Points

- At small data sizes the limit is generally RoI Distributor.
- At larger data sizes the limit is the RoB or the FeX.
- The limiting average time/event in μs for the configurations are:

| n | n-n-n | n-1-1 | n-n-1 |
|---|----------|----------|---------|
| 1 | 42 (39) | 42 (39) | 42 (39) |
| 2 | 41 (51?) | 53 (85) | 66 (90) |
| 3 | 40 (37) | 64 (118) | 93 (-) |

- There is a small but significant reduction in the RoI Dist. Speed as the SCI traffic increases.
- The SCI transfers have $\sim 10 \mu\text{s}$ pauses when many events are allowed in the system. (Due to shared RoB PCI bus)

2RoB 1FeX 1Global 1024 bytes each RoB Nqueue=4



- The SCI switch was added and improved the performance slightly.
- Loading of the SCI ring of $\sim 30\%$ was observed.

- For multiple RoB-FeX-Global stripes we have scaling, the ratios between the average times/events in μs are

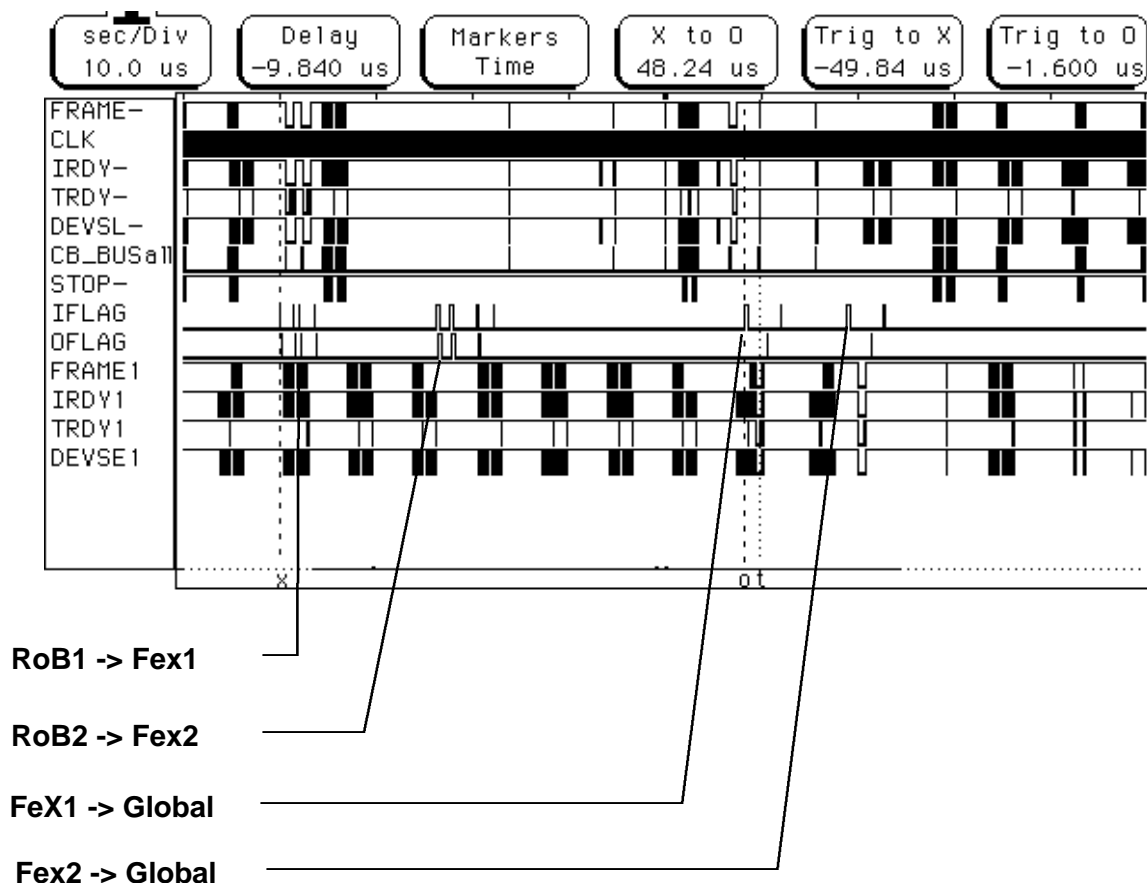
| | 64 Bytes | 1024 Bytes |
|--------|----------|------------|
| 1R1F1G | 1.0 | 1.0 |
| 2R2F2G | 2.03 | 1.9 |
| 3R3F3G | 2.8 | 2.8 |

RoI Dist. / Crate Working harder - queuing is early in system
 For event size > 512 bytes, starting to push the Supervisor,
 no. events peaks bin 23 not 24

- Limiting Aggregate RoB to FeX data rate is:

| No. Stripes | MB/s DMA mode | MB/s Trans mode |
|-------------|---------------|-----------------|
| 1 | 20 | 18 |
| 2 | 34 | - |
| 3 | 42 | ~60 |

- For longer data transfers, the increase in the time of RoB+FeX is that of the slower element.
- For fragment building, it is the slowest chain that counts.
- Less parallelism seen in the SCI transfers than expected for multiple RoBs to FeX. Often the RoI Dist sequence time.



Summary

- Max closed loop event rate achieved = 27 kHz
- Supervisor limit in loopback was 33 kHz
- Highest data transfer rates achieved - would have been higher if not limited by VME transfers in the system.
- Found the data transfers more serial than expected - due to shared components and use of VME.
- Need for care when implementing with shared components/buses.
- The SCI network was stable and robust and did not require data compression.
- The SCI hardware protocols ensure equal access for all nodes and forward progress of the data.
- The interfaces and SCI network can operate at the message level - the software does not have to worry about packet loss.
- Both DMA and Transparent modes have advantages - use of an optimum mix.
- Error detection and rudimentary recovery were included - and tested during setting up of the system !
- Run control and configuration were essential but need more work on the software - Ideal Pilot project work.
- The Local-Global model was stable and worked well.
- The architecture and Technology met the data rates calculated in the paper model.
- There is confidence that the architecture and Technology would meet the requirements of the ATLAS trigger.
- Many people from many institutes have successfully cooperated and worked together to achieve DemoB