Ethernet for the ATLAS second level trigger?

Bob Dobinson, Stefan Haas and Brian Martin,
CERN, Geneva

&

Marc Dobson, Frank Saka and John Strong,

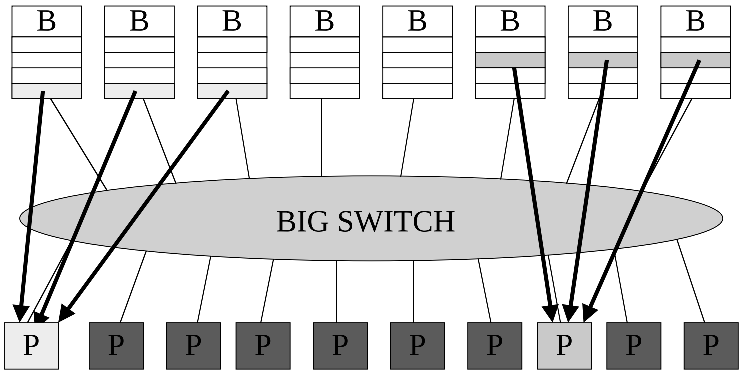
Royal Holloway College, University of London

Talk outline

- ATLAS level 2 trigger requirements
- Results from a 1024 node switching testbed
- The evolution of Ethernet
- Advantages of Ethernet for ATLAS level 2
- Baseline Ethernet measurements
- Ethernet switching fabrics
- The network interface problem
- Conclusions

The Problem: up to 100 kHz images processing rate, 5Gbytes/s data rate

1500 buffers B (distributed 1Mbyte image)



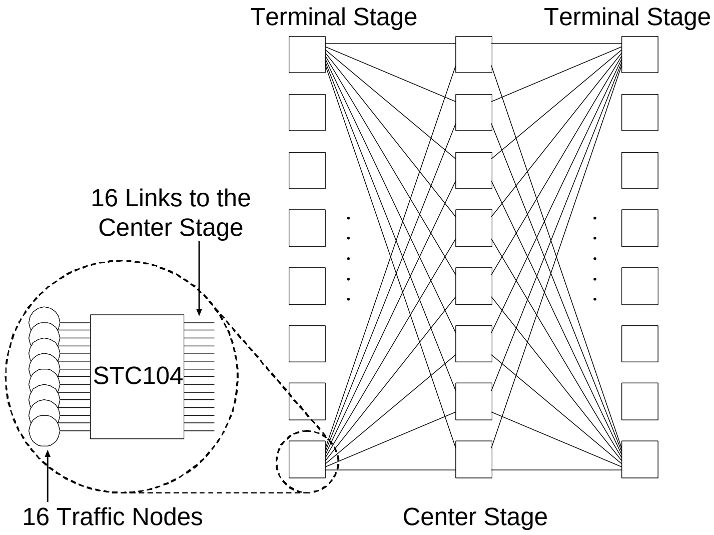
O(1000) 500Mips processors P analyse data from 5% buffers

ATLAS estimated requirements

- Large “scalable” switching fabric, peak throughput in excess of 10 Gbytes/s
- Efficient message passing between network nodes for messages with lengths approximately 100-1000 bytes
- Rates per node
  - Buffers up to 32kHz and 12 Mbytes/s
  - Processors up to 13kHz and 8Mbytes/s

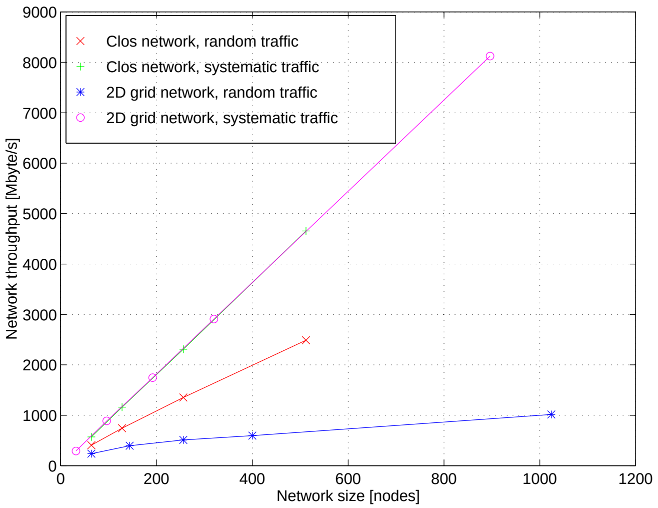
The MACRAME switching testbed

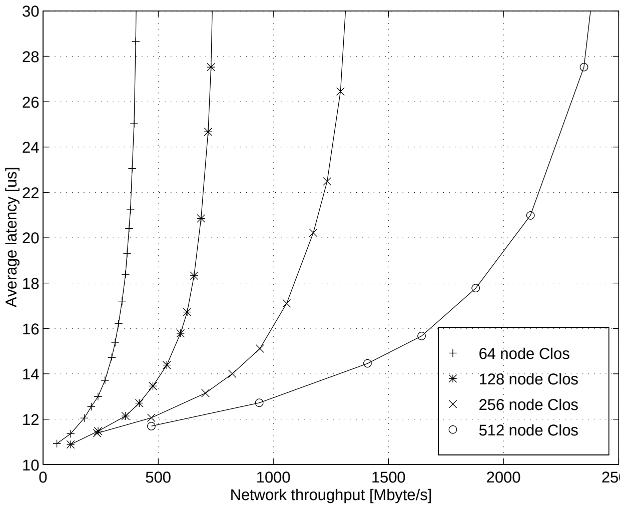
- Very large switching testbed funded by EU
  - Uses 100Mbps DS links and 32 port C104 packet switches
  - Switching fabric is configurable as Clos network, grid, torus, hypercube etc
  - Network nodes can be preloaded with predetermined traffic patterns, packet dispatching overhead only 0.5µs



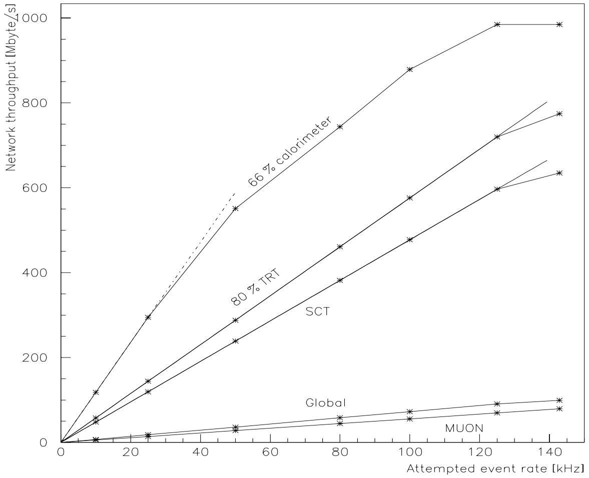
The MACRAME switching testbed

- Measurements
  - throughput and latency as a function of
    - switch topology
    - traffic patterns and rate
      - Random
      - Systematic
      - ATLAS level 2
  - scalability an important issue





Results due to Nina Madsen



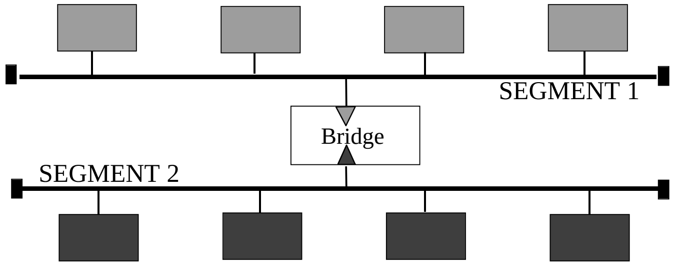
A last word on MACRAME

- The results can be said to represent an upper bound on network performance
- There is essentially no node overhead in dispatching packets, real nodes would behave in a less performant fashion

The evolution of Ethernet

- Originally 10 Mbps CSMA-CD, shared coaxial cable segment, shared bus.
- Half duplex
- Later moved to twisted pair connections to a hub, logically a shared bus still

Ethernet bridges



SEGMENT 1

Bridge

SEGMENT 2

Packets to local destinations remain on local segment
Packets not local passed across bridge
Bridge port learns who is local

Recent developments

- 100 Mbps Fast Ethernet
- Emphasis away from shared segments towards point to point links and switches. Switched 100 Mbps on desk top
- Point to point links allow full duplex operation
- Packet based flow control
- A move towards DS links and switches!

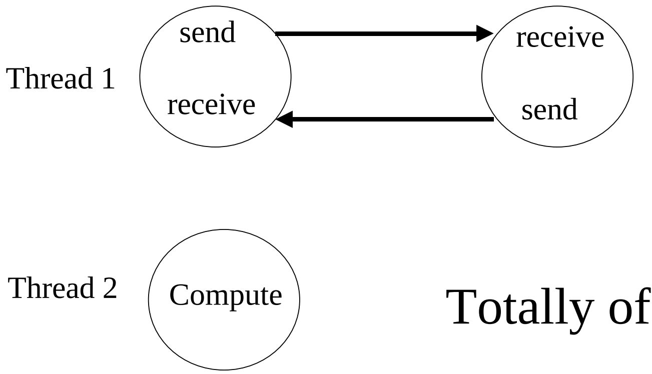
Gigabit developments

- Rapid move from 100 Mbps to a new Gigabit Ethernet standard
- Products available now; network interfaces, switches, testers etc
- Seen as a backbone interconnect but people predict it will end up on the desk-top too.

Advantages of Ethernet for ATLAS level 2 trigger

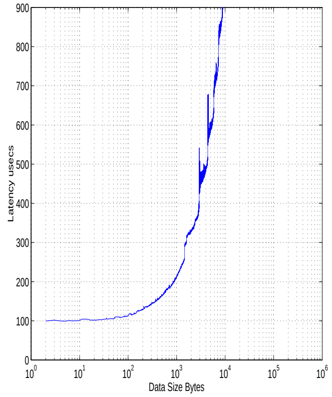
- Huge installed base, unlikely to be displaced as the commodity interconnect
- Highly competitive market, low prices.
- LHC start up 2005, lifetime of equipment in excess of decade. Ethernet will be around!
- Natural to ask “can it do the job”
- Combined with commodity PCs gives an off-the-shelf approach

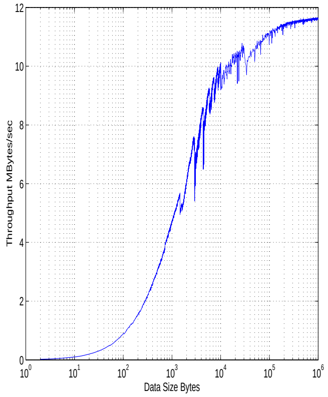
Base line measurements



2x 200 MHz
Pentium
Ethernet Express
Pro 100 NIC
LINUX V4.2
TCP/IP sockets

Totally off the shelf





Summary of results

$t_{elapsed}$ = elapsed time sender to receiver

= t_{zero} + message length / R_{assym}

R_{assym} = asymptotic data rate = 11.6 Mbytes/s

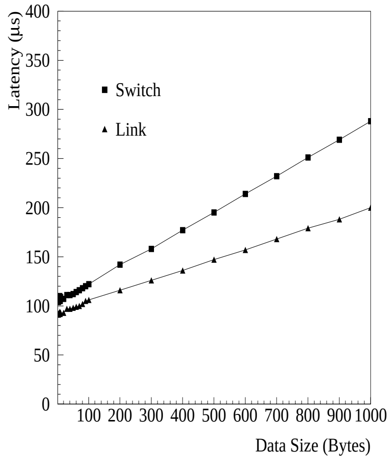
t_{zero} = fixed overhead for zero length message=100µs

t_{CPU} = average CPU time for send or receive = 40µs

Implications for ATLAS

- 40 µs CPU overhead on 200MHz Pentium
  - Would use 130% of the CPU communicating at 32kHz
  - Limit the data transfer rate for ATLAS size messages to well below 12 Mbytes/s
- CPU power increasing x 2 every 18 months, for constant architecture, the overhead should decrease as clock speed increases
- But more powerful CPU consumes more data → more messages

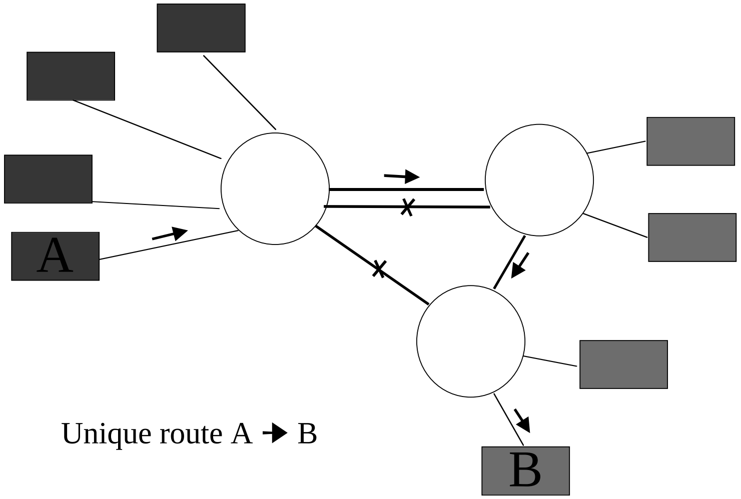
Store and forward switches



- Delay through switch is one packet time. For minimum packet length, about 6µs, we measured 13µs
- Switch delay increase linearly with packet length

Building Large Ethernet Switching Fabrics

- Problem 1.
  - Normally commercial switches dynamically learn the required routing between sources and destination
  - This imposes topology constraints → no loops → only a single connection between switching elements
    - Limits overall bandwidth through switch fabric



Unique route A → B

Solutions

- Disable learning (learning not essential) → load static routing tables
- Use higher speed inter-switch connections
  - Gigabit link between 100 Mbps switch elements
- Treat several physical connections as one logical connection (various manufacturer specific implementations)

Building Large Ethernet Switching Fabrics

- Problem 2
  - The use of store and forward Ethernet switches to build multi-stage networks will increase the latency considerably

Solution

- Industry offers cut through switches → routing once header has been looked at.
- But store and forward still necessary when packets traverse link speed boundary (e.g.100 Mbps to 1 Gbps)
- Learn to live with long latencies, size of buffers B increases but memory is cheap

Network interface issues, reducing the overhead

- The mechanisms are well known
  - Overlap communication and computation
  - Minimise interrupts
  - Avoid memory to memory copies
  - Avoid operating system calls and context switches
  - Implement light weight protocols and simple API
- Dealt with by smart NIC and SW design

Latency hiding

- Latency to fetch data through Ethernet switching fabrics may be long, hundreds of µs
- However, as long as the processors can be kept busy treating multiple events this may not matter
  - Requires low context switching multiprocessing kernel (helped by a smart NIC)

Conclusion

- Ethernet is an option well worth exploring
  - ATLAS level 2 trigger pilot project will address this issue over the next two years