IEEE802.3 Ethernet
Current Status and Prospects for LHC

R. Dobinson  CERN
M. Dobson    CERN
S. Haas      CERN
B. Martin    CERN
M.J. LeVine  CERN, Brookhaven National Laboratory
F. Saka      CERN, Royal Holloway College
Smoke and Mirrors in Lyon

What has this ..

...got to do with LHC?
The 10 Mbps Ethernet Standard
Been there and still doing that!

• First defined as a shared resource of coaxial cable
  – Carrier Sense Multiple Access/Collision Detect
  – Half duplex
• First lab implementation in 1973, published in 1976
• First controller boards available in 1982
• IEEE 802.3 first released in 1983

It’s still wired to my office today
Whatever happened to Moore’s law?
Architectural changes

- Cable segment is limited
  - length and load and management
- 1990 10 Base T
  - Unshielded Twisted pair
  - Hub based Star topology
  - Easier to manage
- Scaling limited by broadcasts
  - Bridges split segments, system can grow
  - Still only half duplex
- Still only 10Mbps /shared segment
Fast Ethernet

- 1995  100Mbs standard: a quantum leap forward
  - Point to point switched over fiber and copper 100BaseFX
- 1997  Full duplex operation for both 10 and 100Mbps
- Can now avoid collision domains, get some distance
- There was a brief hope for cut through switches
  - Can’t do that with mixed 100M & 10 M segments
- Market stayed with store and forward
- Flow control using pause frames
  - no arbitrarily dropped frames
  - sounds good but in fact it’s only partially true
1998 Gigabit Ethernet

- It kept the CSMA_CD
  - But nobody uses it

<table>
<thead>
<tr>
<th></th>
<th>Ethernet</th>
<th>Fast Ethernet</th>
<th>Gigabit Ethernet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Rate</td>
<td>10 Mbps</td>
<td>100 Mbps</td>
<td>1000 Mbps</td>
</tr>
<tr>
<td>Cat 5 UTP</td>
<td>100 m</td>
<td>100 m</td>
<td>100 m</td>
</tr>
<tr>
<td>STP/Coax</td>
<td>500 m</td>
<td>100 m</td>
<td>25 m</td>
</tr>
<tr>
<td>Multimode Fiber</td>
<td>2 km</td>
<td>412 m</td>
<td>550 m</td>
</tr>
<tr>
<td>Single-mode Fiber</td>
<td>25 km</td>
<td>20 km</td>
<td>5 km</td>
</tr>
</tbody>
</table>
Commodity Volume and Prices

Number of ports produced

Cost per port
What have we got today: Connectivity to go!

Today we have options to build networks with

– Point to point link and switch technology
– Choice of media
– Distances of kilometers
– Up to 1Gbps to and from every processor
– Commodity prices

But is it enough for DAQ applications?

And where’s the dirt?
Presentations today from the four LHC experiments. First pass analysis indicates that with a mixture of 100M and 1G networks……. “It looks about right”

So what’s to worry about?
Worry #1
The NIC bottleneck

• The TCP/IP heritage
  – it works very well
  – and quite fast too
  – if you give it
  – all your CPU!

– PCI bus (33Mhz/32 bits) is a limit at 1Gbps
What’s to be done?

• Option 1: Lightweight protocols
  – lose out on automatic error recovery, need a strategy
  – Still limited by the NIC and/or PCI

• Option 2: use programmable NIC’s
  – Exports TCP/IP or other protocol to the NIC
  – Still limited by PCI and OS scheduling

• Option 3: Combine 1 and 2
  – Network processor <-> user space I/O
  – Still limited by PCI -> buy faster PCI
Lightweight protocol versus UDP at 100Mbps and 1000Mbps

Streaming on 400Mhz PC
At 1Gbps
Neither reaches full BW
Lightweight 45MB/s
UDP 70MB/s

at 100Mbps full BW
Lightweight 100 bytes
UDP 200 bytes
CPU cost of communications

With lightweight protocols
- independent of frame size
- NIC dependence

With TCP/IP
- frame size dependence
- independent of line speed
- never reaches LVL2 rate
Worry#2
Does the switch scale?

- Physics requirements
  - Any buffer to any processor
- Topology requirements
  - Star configuration
- Outer layers here today
- Characterise them, build models and extrapolate
- Extrapolating two orders of magnitude is scary
Worry #3
What’s the box in the middle?

High performance switching
- Physical Layer (PHY)
- Media Access Controller (MAC)
- Network processing (NP)
- Fabric adapter (ADAPT)
- Components up to 2.5 Gbps now
- Some at 10Gbps, others very soon
- Valency (no of ports) & link speed major issues
Worry #4
How big and how fast will it be?

- Single chip ethernet switches ‘now’
  - sampling Q4 2000 (for volume read Q3 2001)
  - 12 @1Gbps+1@10Gbps bandwidth up to 44 Gbps
- Crossbar Switch Chips next year
  - Up to 32 ports @ 2.5 Gbps on one chip
  - Enables Terabit switching fabrics
- Largest switch boxes today 64 ports
- Expect 256 to 512 port boxes in 2 to 3 years
Worry#5  Use it and lose it?

- Generic problem: no guarantee of delivery
- Flow Control only at the link level
- Some vendors propagate flow control
  - causes tree saturation and head of line blocking
- Some just trash the overflows
  - penalises the most popular port
- Variable delivery rates for uni/multi/broadcast
- Not all will guarantee broadcast delivery
- Protective protocols required!
- Need to understand the product
What’s Next

Evolution

Time


Bandwidth 10Gbps

Collision Free
Full Duplex

Bridges

Switches

UTP

My office

Coax Cable

Collisions

Mbits/sec

10000

1000

100

10

1
Techno-Surfing Questions
And your starter for 10 Gbps is:

- When can you buy it
  - the PHY level from Lucent later this afternoon
- When will the standard settle down
  - 2002 but pre-standard kit released well in advance
- When will the price go commodity
  - depends on Napster staying in business
- Industry standard response:

  The answer is Q3 2001, what was your question?
10G Ethernet Objectives

- Define two families of PHYs
  - LAN PHY, 10.000 Gb/s
  - WAN PHY, rate compatible with the payload rate of OC-192c/SDH VC-4-64c

- support link distances of:
  - At least 100 m over installed MMF
  - At least 40 km over SMF
10GE Objectives

• Preserve the 802.3/Ethernet frame format.
• Preserve minimum and maximum FrameSize of current 802.3 Std. (i.e. no jumbo frames)
• Support full-duplex operation only.
• Support star-wired local area networks
1 & 10 Gbps opens up to distance

- Point to point over a leased fiber
  - Carriers more likely to lease a λ

- Point to points over a MAN/WAN
  - Ethernet over Sonet
  - Get resilience and management
  - But not for free

- Point to points as raw Ethernet
  - Bell-Head versus Net-Head
New data network business model

• If you don’t want or need Sonet protection
  – You can run raw ethernet over commercial fiber
• 1G Ethernet ran raw over 1062kms of the MONET ring around Washington
• Montreal: Gigabit Ethernet over same MAN fiber …. which delivers cable TV
• five year contracts today for point to point, 1Gbps data services at a less than US$700 /kilometer/month
• Cost for a 1 Gbps Ethernet data network is 20% more than 50 Mbps managed SONET data services from a carrier (http://www.cookreport.com/09.07.shtml.)
Nobel prize from your armchair?

- LAN to home? already there in Sweden, sort of
  - fiber to the apartment switch in the basement,
  - 10Mb Ethernet to every apartment [http://www.bredband.com](http://www.bredband.com)
  - installation cost ~20% equivalent ADSL
  - aggregate external 5Kbps/user, $25/month, always on
  - More lines increases the local traffic, not external

- An MP3 or Physics event file is about 2Mb
  - 100Mbps to the home = 6 events/sec
  - Distributed Level 3 Event processing?

- Still need to do shifts to get name on the papers
10 Gbps Switching

How is it done? There is smoke

- You can switch 10Gbps packets with Silicon technology
- 2.5 Gbps signaling
  - is easy on a chip,
  - OK between chips,
  - Tricky on a backplane
- optical backplanes to the rescue
- 10 Gig switches compact DAQ architectures
  - No-one wants to pull 1000+ cables to a switch rack
And There are Mirrors

- Optical circuit switching is already here
- Electronic controlled optical packet switching
  - still in the lab
  - debated for 10Gig ->almost inevitable at 40Gig
- Terabits on the fiber are easily transported
  - But it’s Terabucks to put them there today
  - 600K lasers in 2000 maybe 2.5M by 2003
- All optical is 10 years away,
  - No optical buffering & label changing a bit tricky
- Components are large, difficult to scale
  - Investors love them
Techno Surf’s up

• Is there a market pull for all this bandwidth?  
  – certainly in the business sector

• Broadband to and from the home?  
  – competition with wireless broadband  
  – cable/ADSL offer asymmetric services

• But VDSL is symmetric, 13Mb/s @ 1km  
  – projected 4.5 million lines by 2003  
  – upstream can be ATM or Ethernet

• ADSL/VDSL market 9 Million lines by 2003
Crystal Ball Forecast
What’s available for the LHC Startup

• Gigabit Ethernet connectivity at the desktop,
  – $250 connection price NIC + port
• The CPU is yours again, communications load on the NIC
• At least 10Mbps to the home, $25/month, ‘always on’
• Ethernet kit everywhere in DAQ, except the front end
• Public networks too unreliable for LHC use
  – Rental fibers or wavelengths between institutes
• Significant On-Line capacity in member states centers