The LHCb Event Building Strategy

Niko Neufeld
CERN, EP Division
Geneva, Switzerland

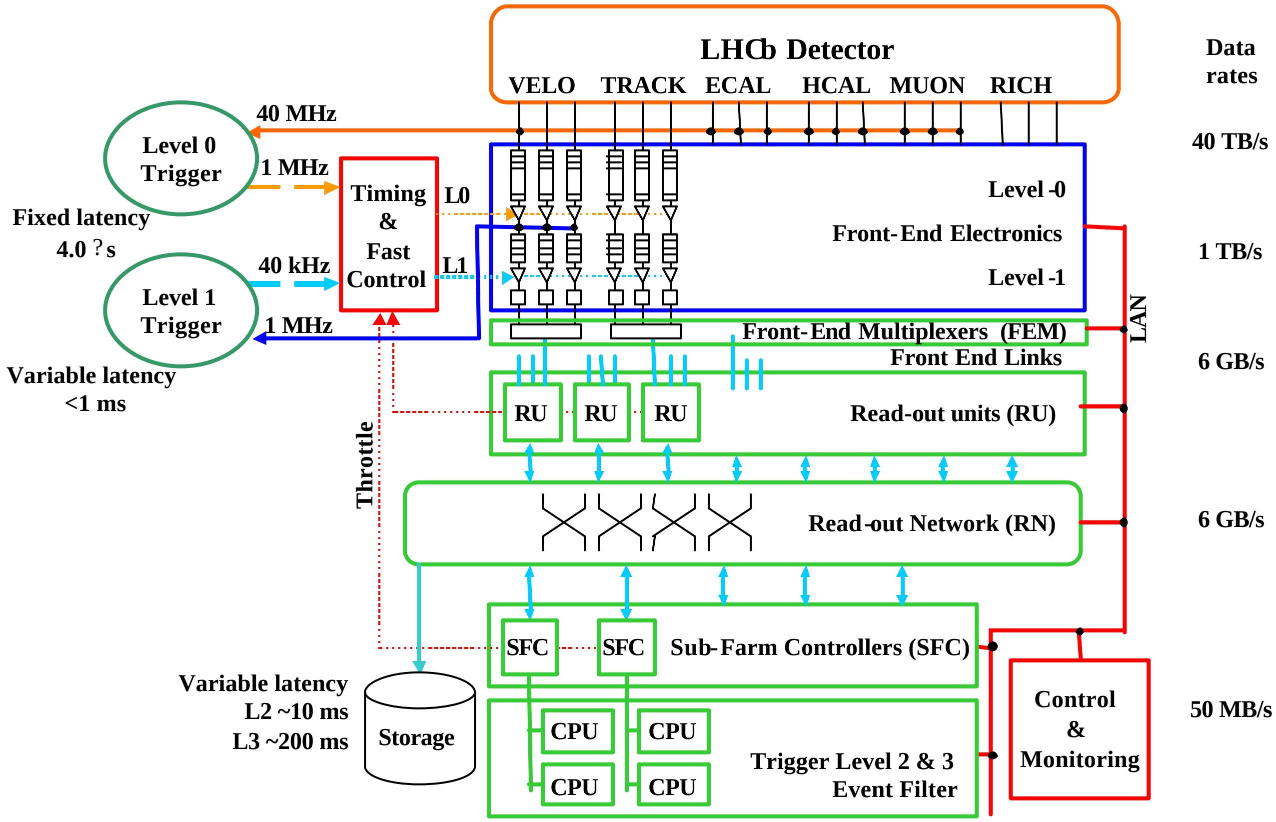
Presentation at IEEE-NPSS Real Time 2001
June 4-8, 2001, Valencia, Spain

Overview

- Architecture of the LHCb DAQ
  - Trigger rates, event size
  - Event building requirements
- Gigabit Ethernet for the Readout Network
- Network topology
  - Commercial switches
  - Small modules

Main Architecture

- Data are **push**ed **through** from the Front-end links to the CPU farm
- No upwards communication
- Throttle to disable trigger in case of persisting contention
- Backpressure (Flow Control) to deal with local contention



Event Building

Event Building consists of two main tasks:

- The fragments of an event, originating from *many* sources must be transported to *one* destination (through a network/bus)
- The fragments must be *arranged in* the *correct order* as a contiguous event
  - Using general purpose or dedicated CPUs such as High End PCs, Network Processors, Smart NICs

Readout Network

- Most likely choice for the Network Technology: Gigabit Ethernet
- Also studied: Myrinet
- Readout Network will be a rather large (~ 128 x 128) Switching Network

- Must sustain at least 40 kHz of fragments ~ 1000 Bytes
- Should provide enough margin to increase input rate to 100 kHz

Implementation of Gigabit Ethernet Switching Network for Event Building

- Conventional:
  - Large Campus/MAN switches (e.g. Foundry Big Iron 120 Gigabit Ethernet ports)
- Alternative:
  - Re-use of NP-based DAQ modules (→ Presentation B. Jost)
  - Basic building block is a 4x4 programmable switch, giving full control and maximum flexibility (in particular for flow control)

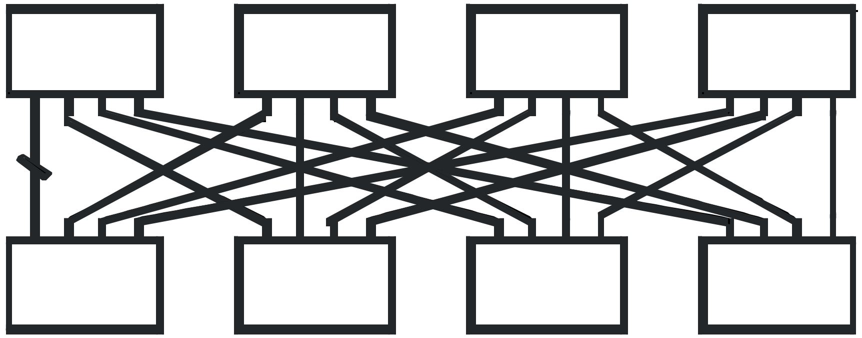
Network Topology: The 2 crucial questions

- How to build a 128 x 128 network out of building blocks with n x n inputs:
  - when n is small, e.g.: 4
  - when n is big, e.g.: ~ 60
- How to optimise the usage of the installed bandwidth, taking into account the direction of the dataflow in the DAQ system

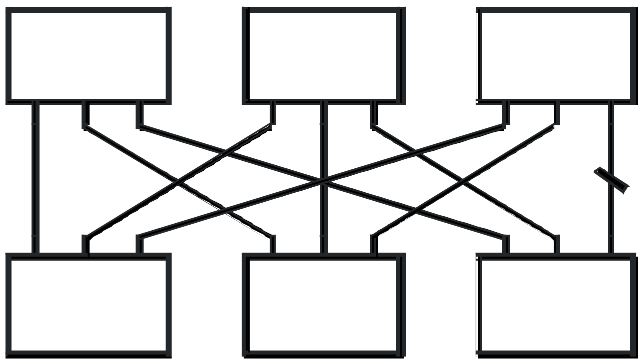
Banyan Network From Large Switches

For a rate of 100 kHz

Max load on single link	60 MB/s (50%)	84 MB/s (70%)
# of input- or output links	240	180
Maximum fragment size	625	875

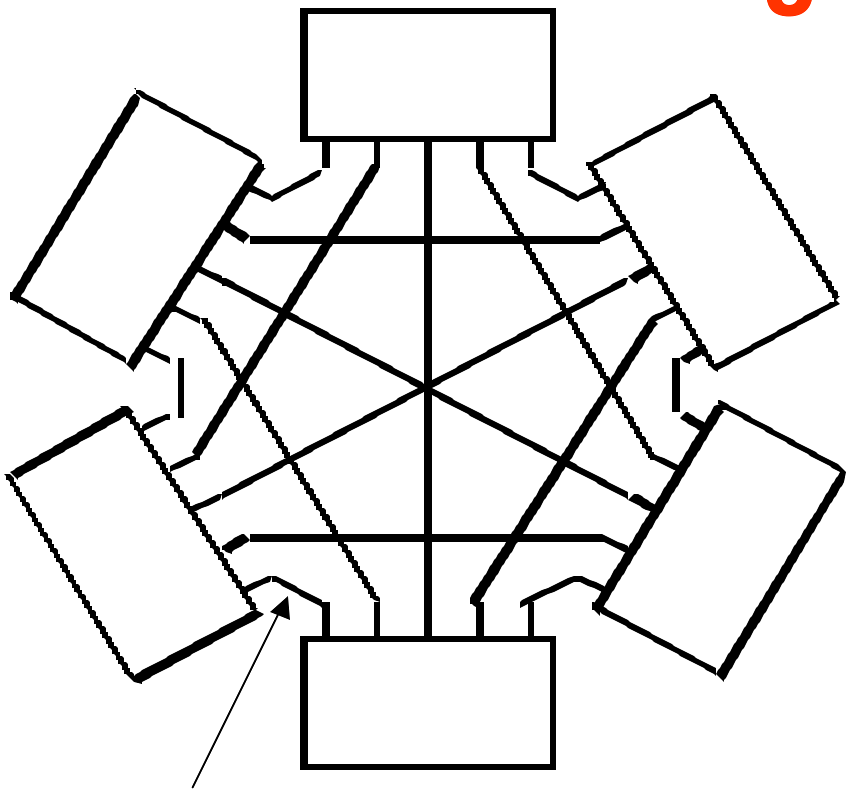


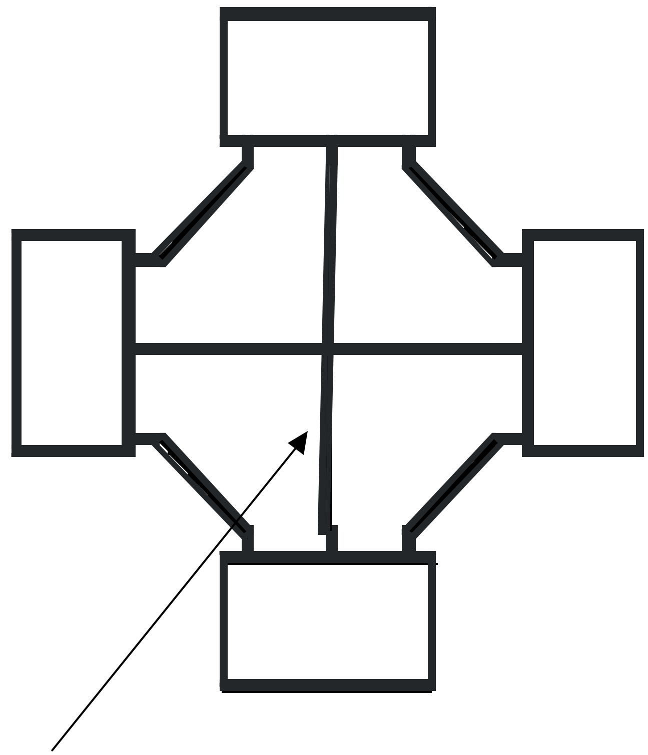
15 links per connection



20 links per connection

Optimised Network From Large Switches



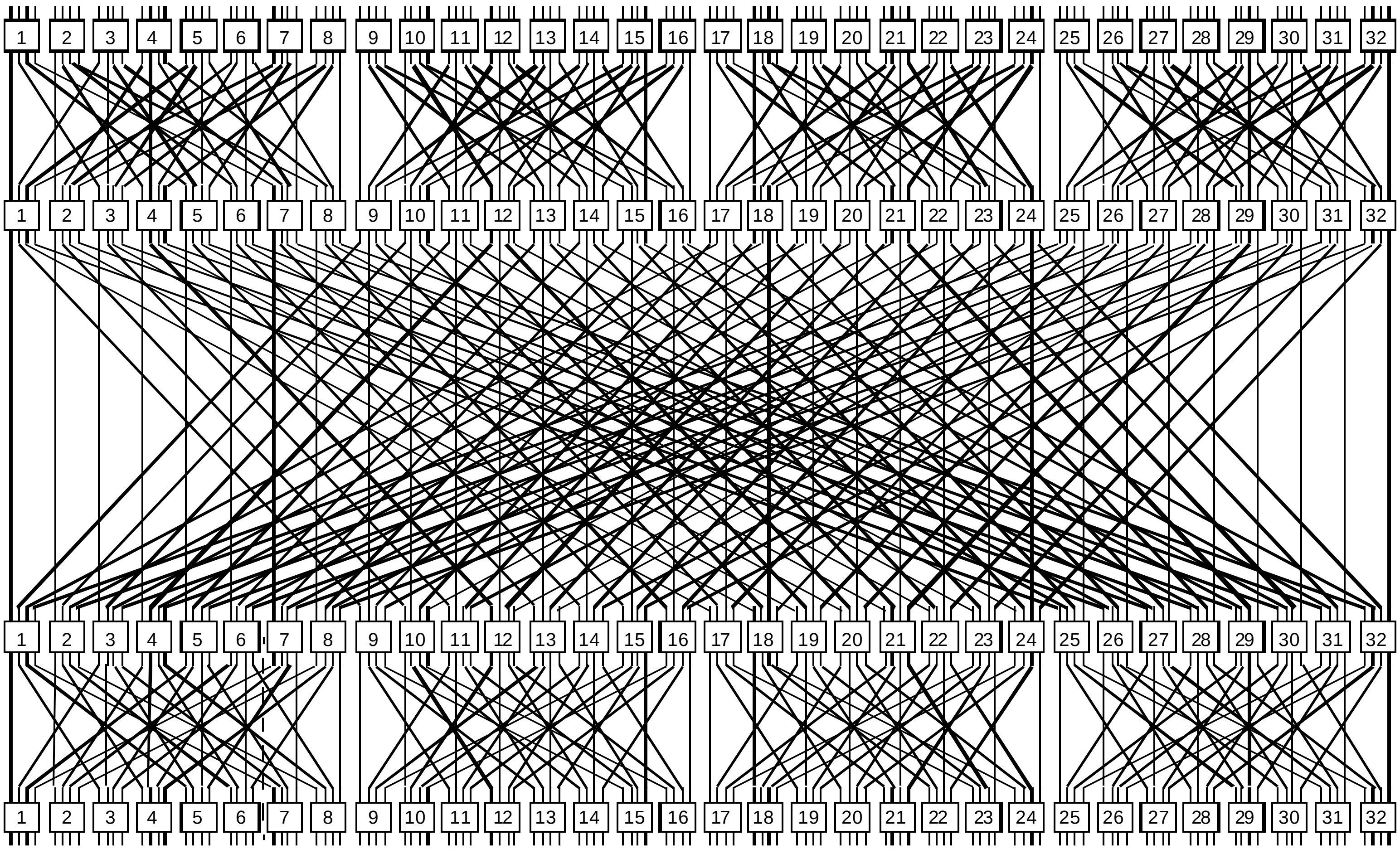


8 links per connection
240 x 240 ports; effective load 40 % @100 kHz

11 links per connection
174 x 174 ports; effective load 72 % @100 kHz

Banyan Network for 4x4 Modules

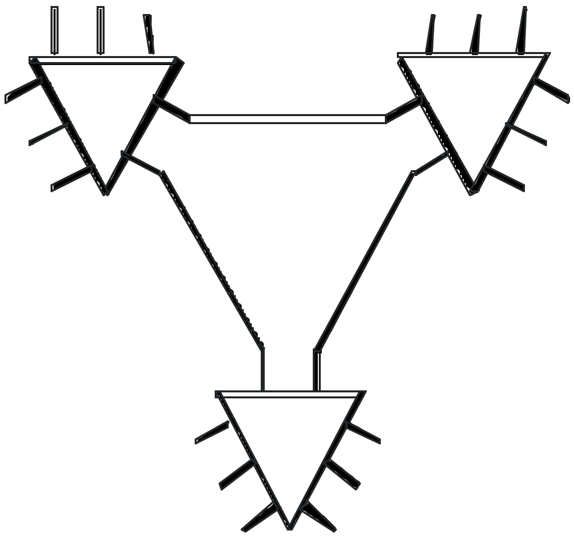
128 X 128 complete connexion based on 32 X 32 sub-switches



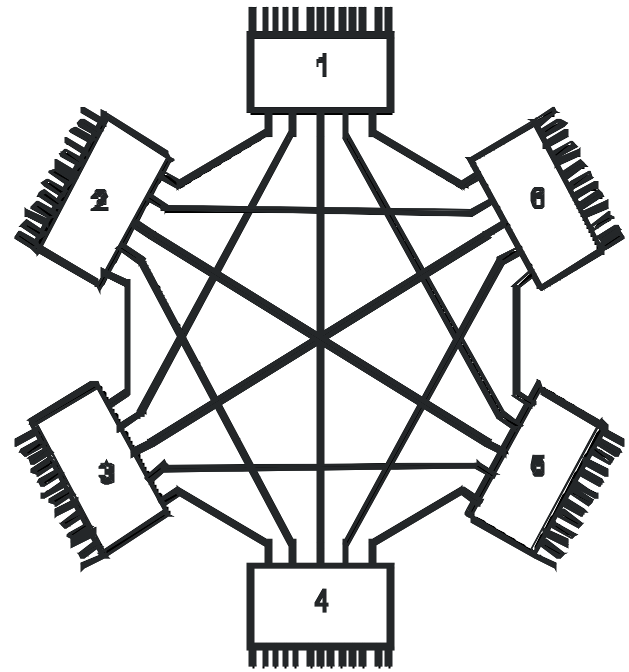
40 kHz:
40% load on input – 39% load on internal inks
128 Modules needed

100 kHz:
50% load on input- 49% load on internal links
256(!) modules needed

Alternative Topology for 4x4 Modules



x 6 =

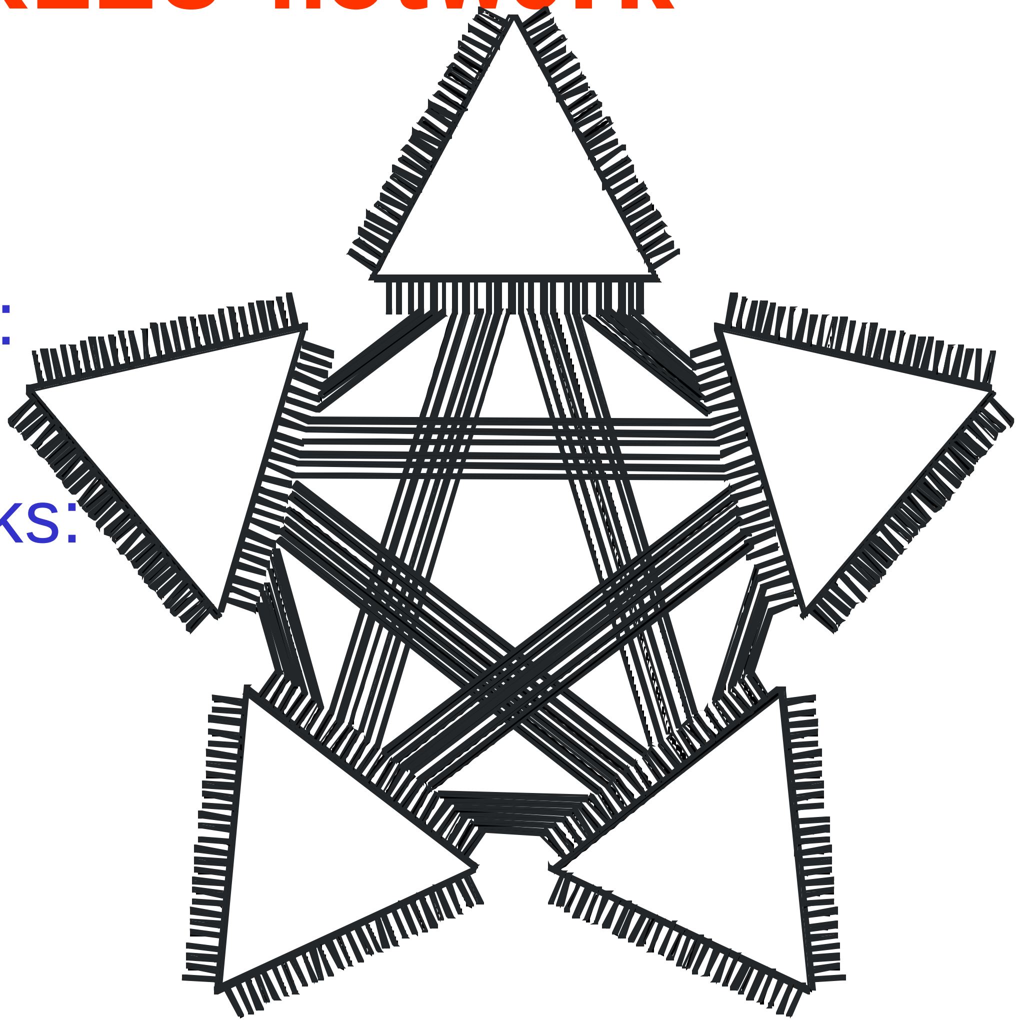


3 modules fully connected make a 9x9 module

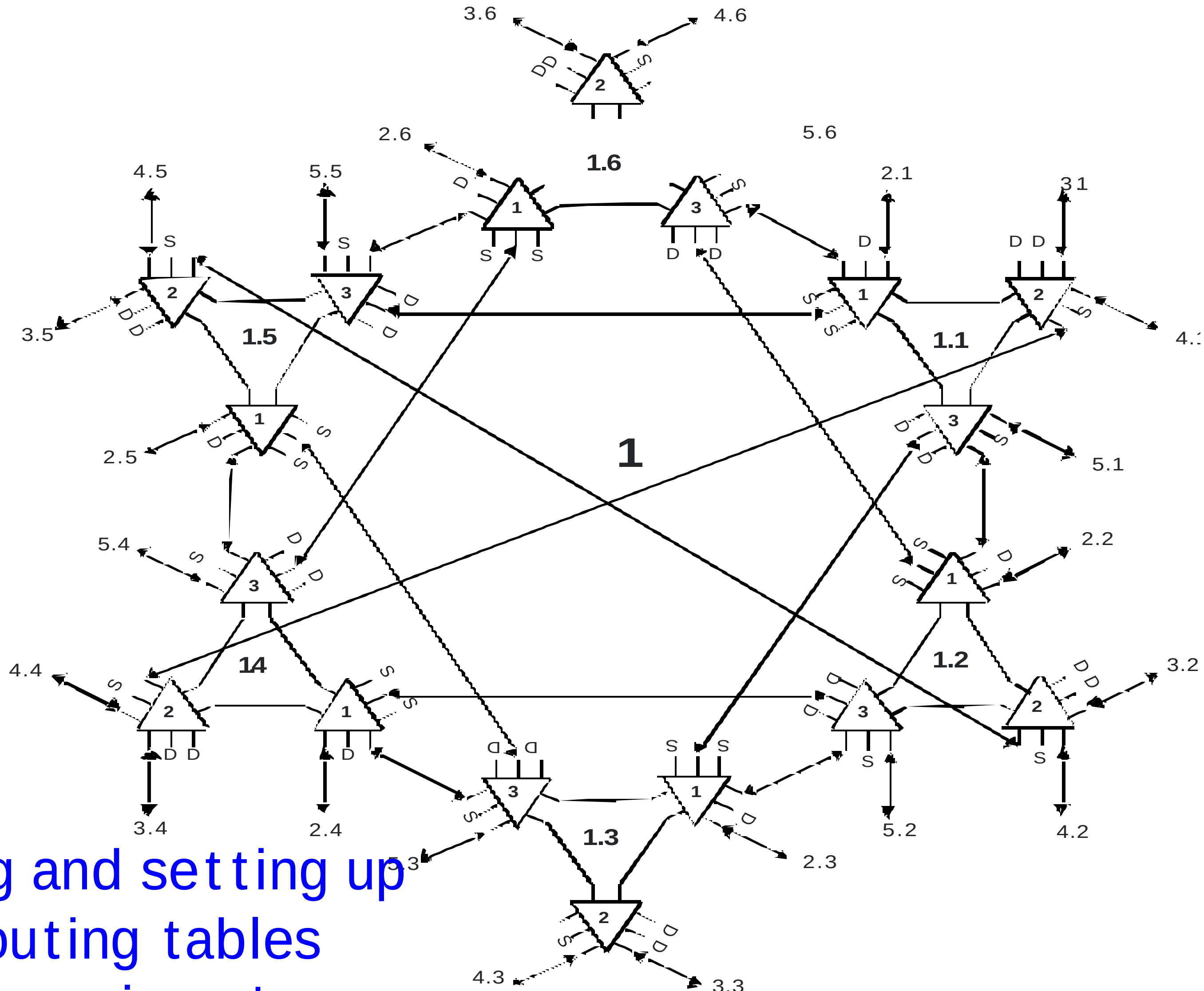
6 modules fully connected make a 39 x 39 module

Fully connected 125x128 network

- Consists of five 39x39 modules
- Load on input ports: 40% @ 40 kHz
- Load on internal links: 34% @ 40 kHz
- **90 4x4 modules needed in total**



Nitty Gritty Connectivity



Cabling and setting up the routing tables become an issue!

Conclusions

- The LHCb Event Building will be done using a Gigabit Ethernet switching network
- Event fragments will flow freely from the front-end links to the entry points of a CPU farm, without synchronisation
- The switching network has a considerable size and cost
- Optimised network topologies can take optimal advantage of the unidirectional data-flow