High performance GbE switches for Data Acquisition Systems

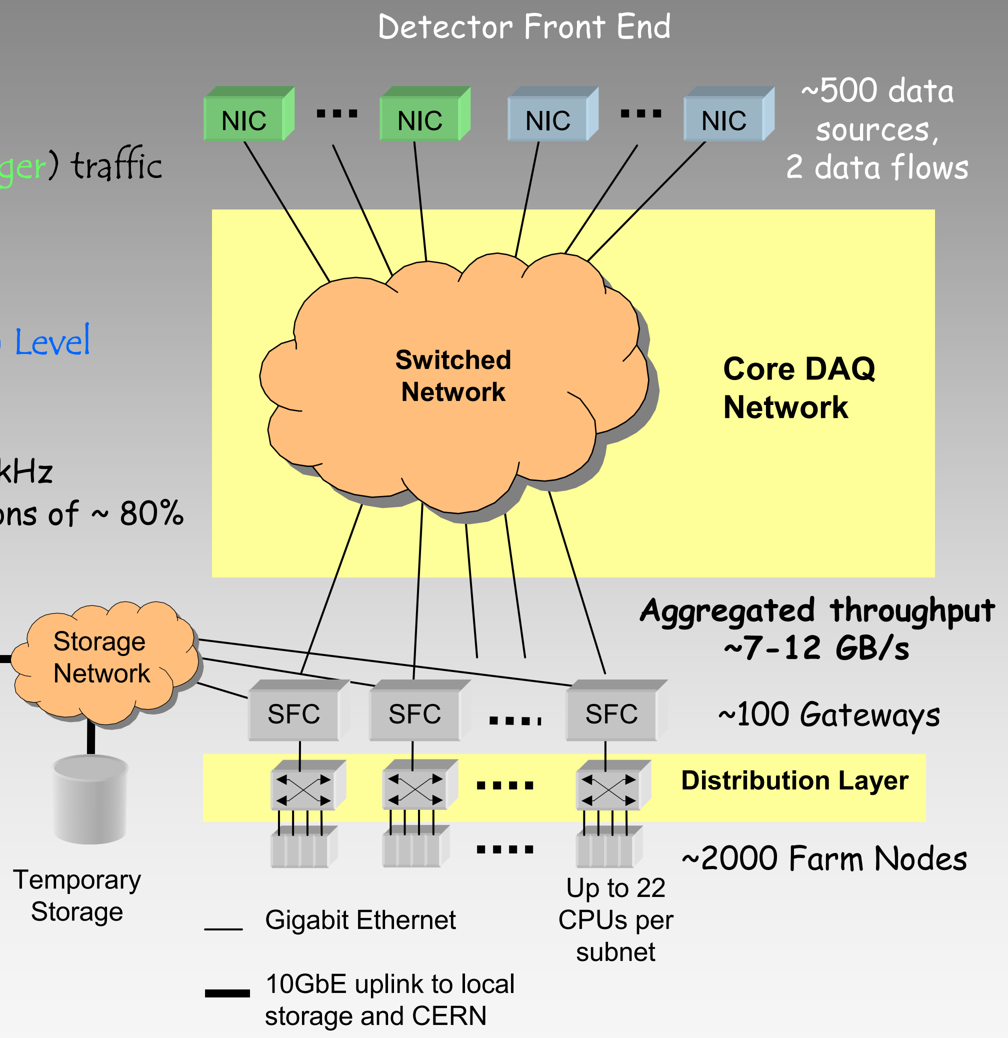
A. Barczyk, J-P. Dufey
for the LHCb collaboration

Overview

- The context: LHCb readout network
- Readout network topology
- Evaluation: LHCb DAQ test-bed
- Simulation: Extrapolation to complete system

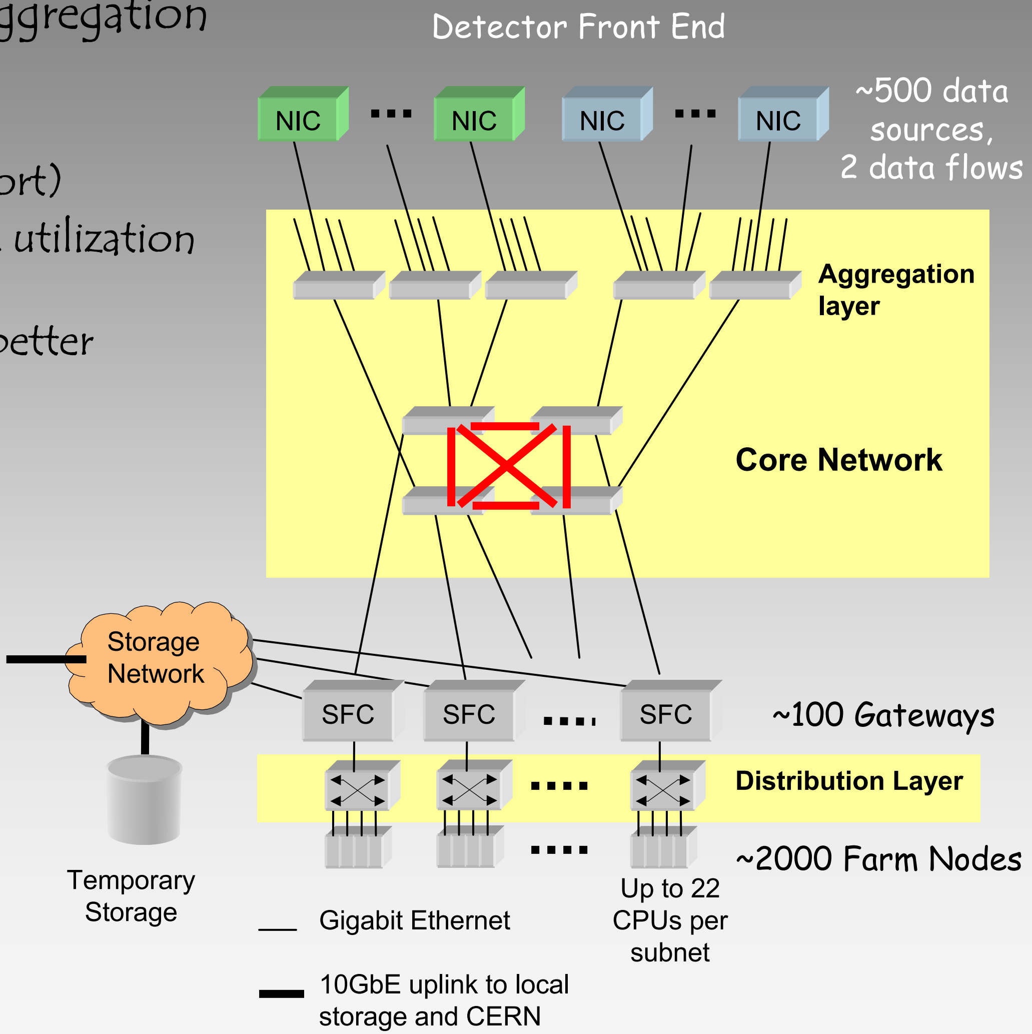
LHCb readout network

- The LHCb readout network is built on Gigabit Ethernet technology
- From network point of view:
  - 120 sources of high priority (Level 1 trigger) traffic
    - Latency constrained
    - Fixed arrival times ~40 kHz
    - ~ 30% link utilization
  - 300 sources of low priority traffic (High Level Trigger)
    - No latency constraints
    - Variable arrival times, mean rate ~4kHz
    - Link utilization 3-30%, with exceptions of ~ 80%
  - ~100 destinations
    - Sub-Farm Controller PC
    - Act as gateways to the CPU farm
    - Perform last stage of event building and distribution to worker nodes
- Event building traffic: all sources contain fragments of the same event → all send data to the same destination (round robin)
- Push protocol throughout
- No data retransmission



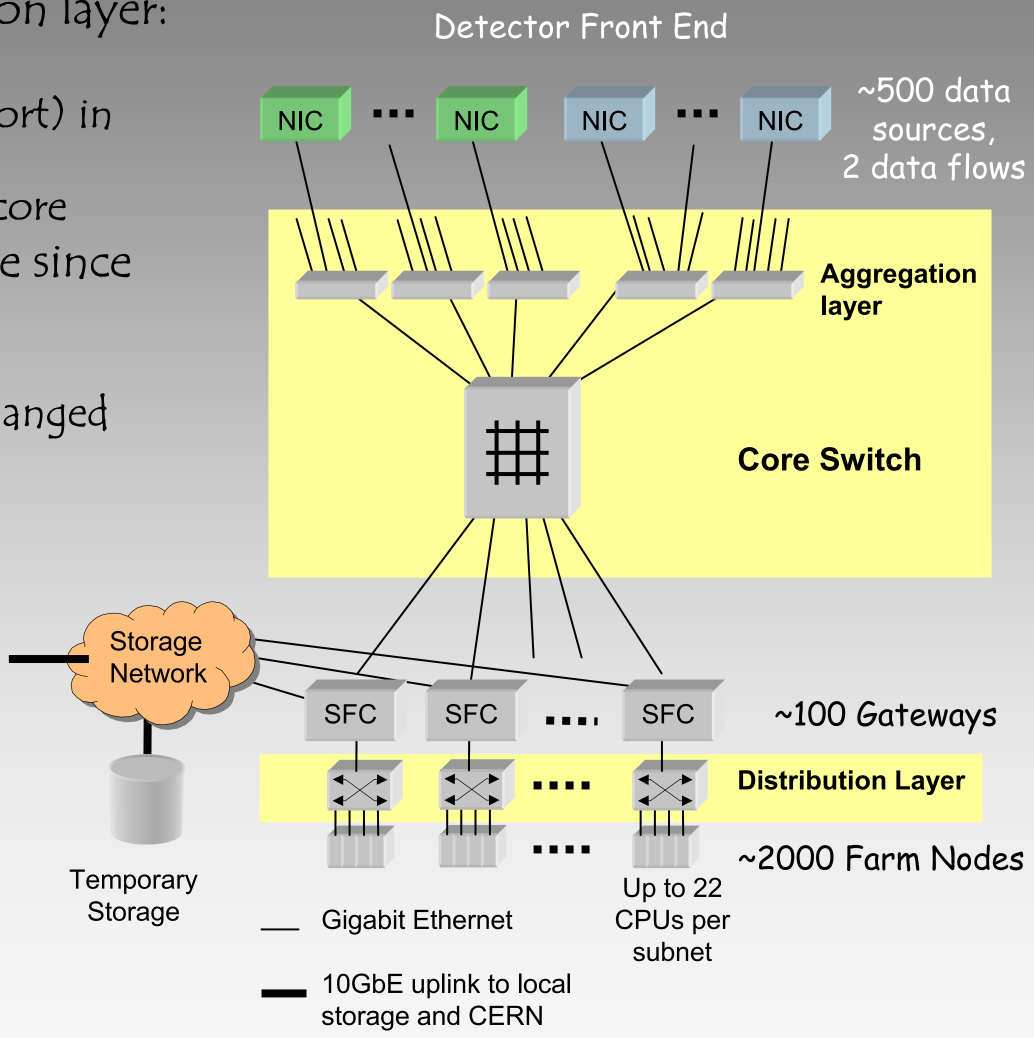
Possible Topologies

- Fully interconnected core with aggregation layer:
  - Multi stage network
  - Low port density switches (48 port)
  - Aggregation layer enhances link utilization into the core
  - Full mesh topology in core for better bandwidth utilization
- Aggregation (edge) switches are
  - Cheap
  - Commodity hardware
- Requires
  - Enough buffer memory
  - Many interconnecting links
  - Link aggregation at all stages (edge → core, mesh)
- Possible use of 10G Ethernet between edge and core
  - Optical → expensive!



Possible Topologies, cont.

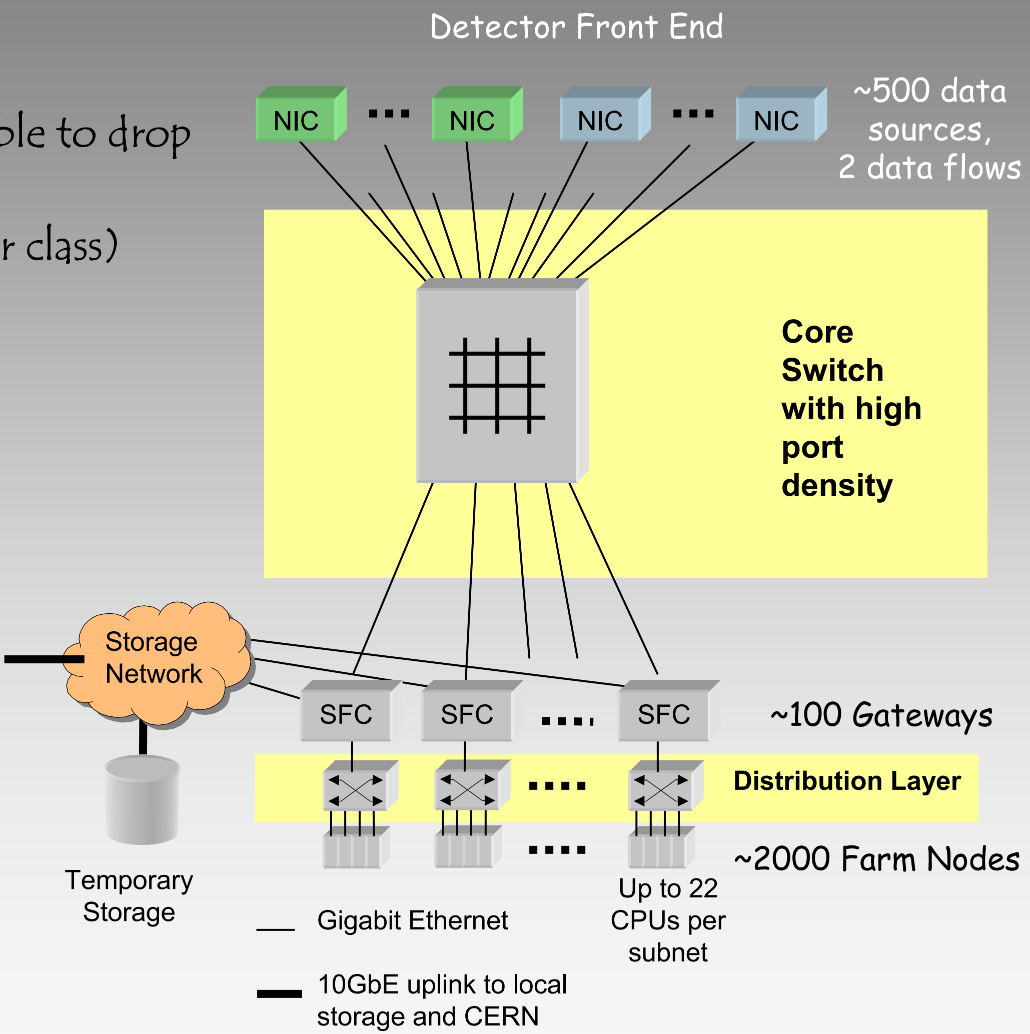
- Single core switch with aggregation layer:
  - Multi stage network
  - Low port density switches (48 port) in aggregation layer
  - Single switch (~ 200) ports in core
- Core switches of this size available since a few years
  - Simpler design
  - Aggregation layer remains unchanged (larger part of the setup)
  - Higher per port cost



- Possible use of 10G Ethernet between edge and core
  - Optical → expensive!

Possible Topologies, cont.

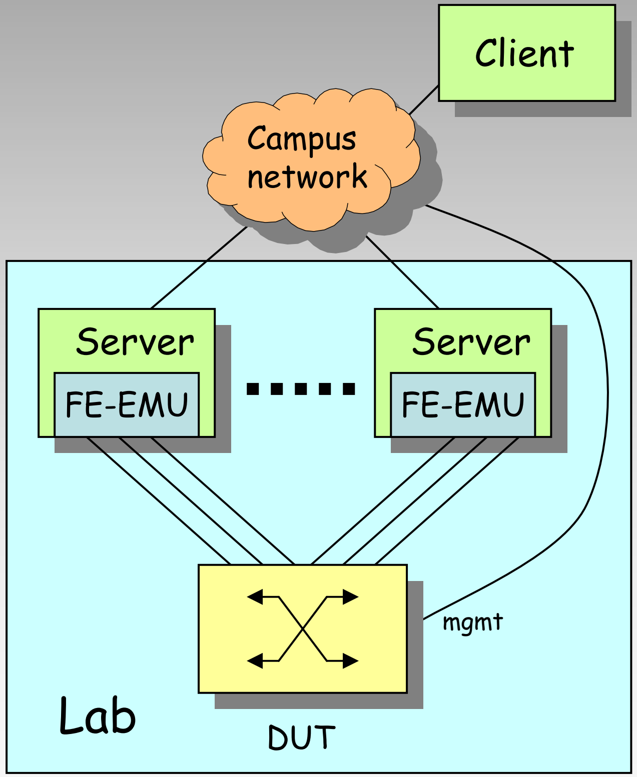
- Single switch core
  - A high port density switch with > 500 ports would make it possible to drop the aggregation layer
  - High performance switch (router class)
  - Higher per port cost
  - Only recently available
- Simpler setup
  - No interconnecting links
  - No link aggregation necessary
  - Simplifies management and performance monitoring



Switch evaluation

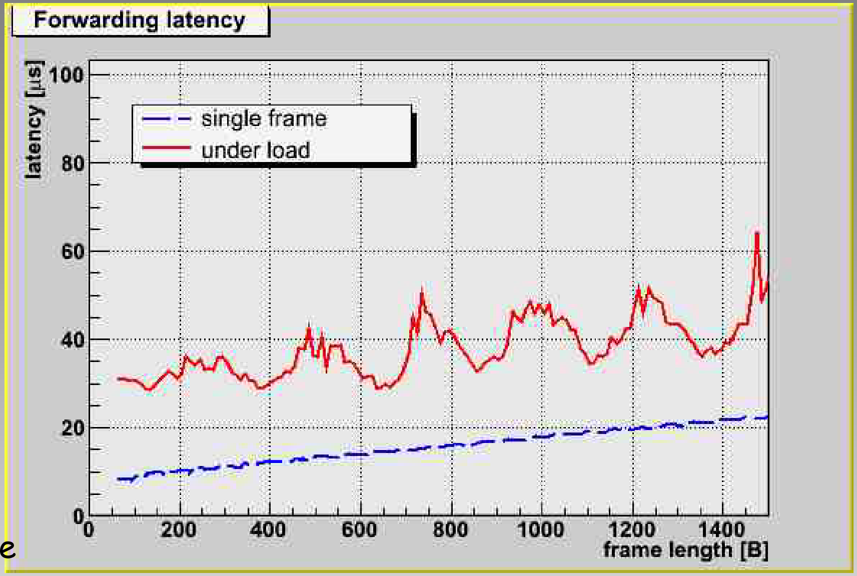
- Parameters we need to measure:
  - Switching latency
  - Egress queue depths
  - Behaviour under LHCb traffic
  - Generic performance tests (full mesh, large statistics packet loss rate, …)
- LHCb DAQ Test-bed:
  - FEE emulators
    - Network Processor based
    - 3 GbE ports per PCI card
    - Fully programmable traffic generators
    - Used also to analyse traffic
  - Client-server application
    - Server running on hosts containing NPs
    - Client running on desktop box
    - Python scripts running tests
      - Downloading test application to NP
      - Defining traffic pattern
  - Test-bed limitations
    - Size: only up to 48 GbE ports available

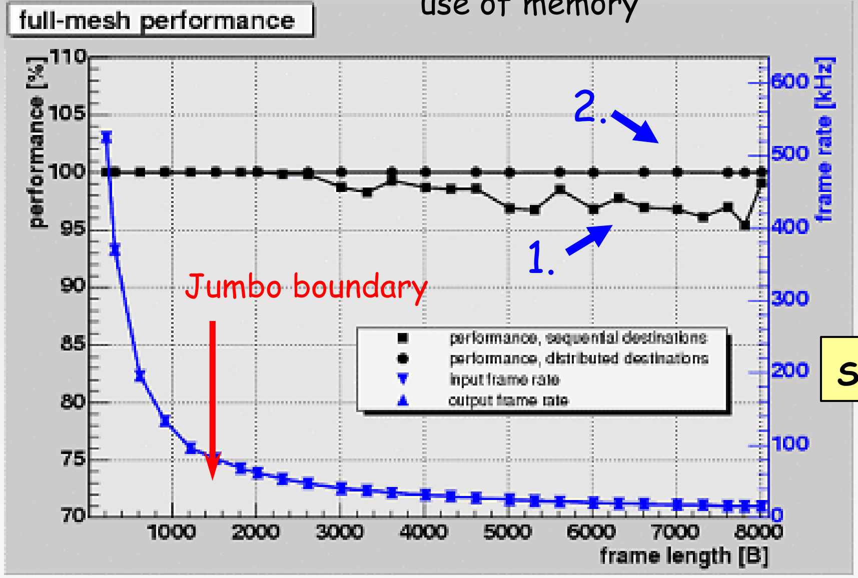
    → use simulation to extrapolate to full-sized system

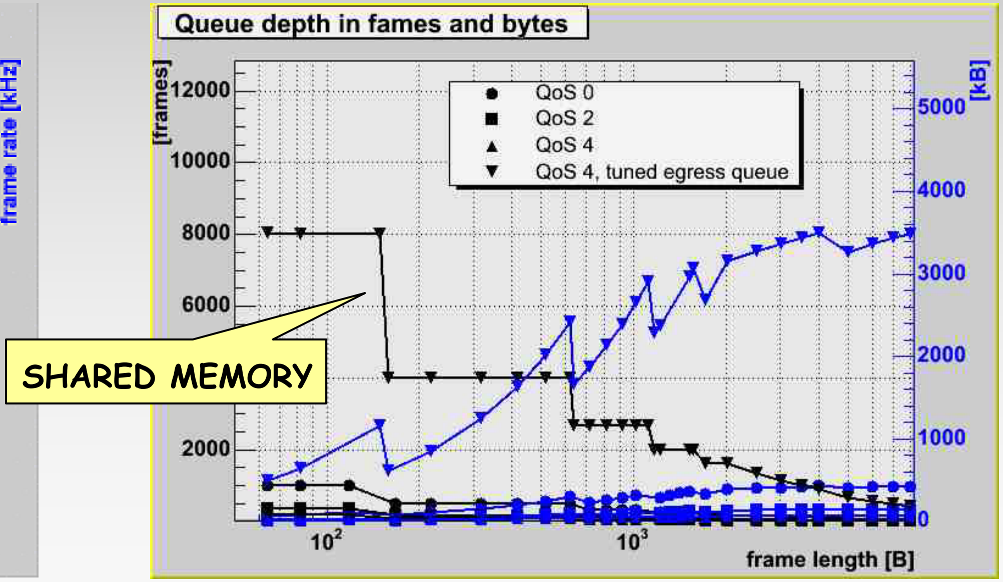


Switch evaluation, cont.

- Latency
  - parameter for simulation
- Queue depths
  - Parameter for simulation
  - Minimum requirements to be met
- Full-mesh performance
  - Scan over frame size
  - Take memory layout into consideration, e.g.:
    1. Consecutive ports might use same memory block → overflow in full-mesh traffic
    2. Jumping by N ports will make more effective use of memory





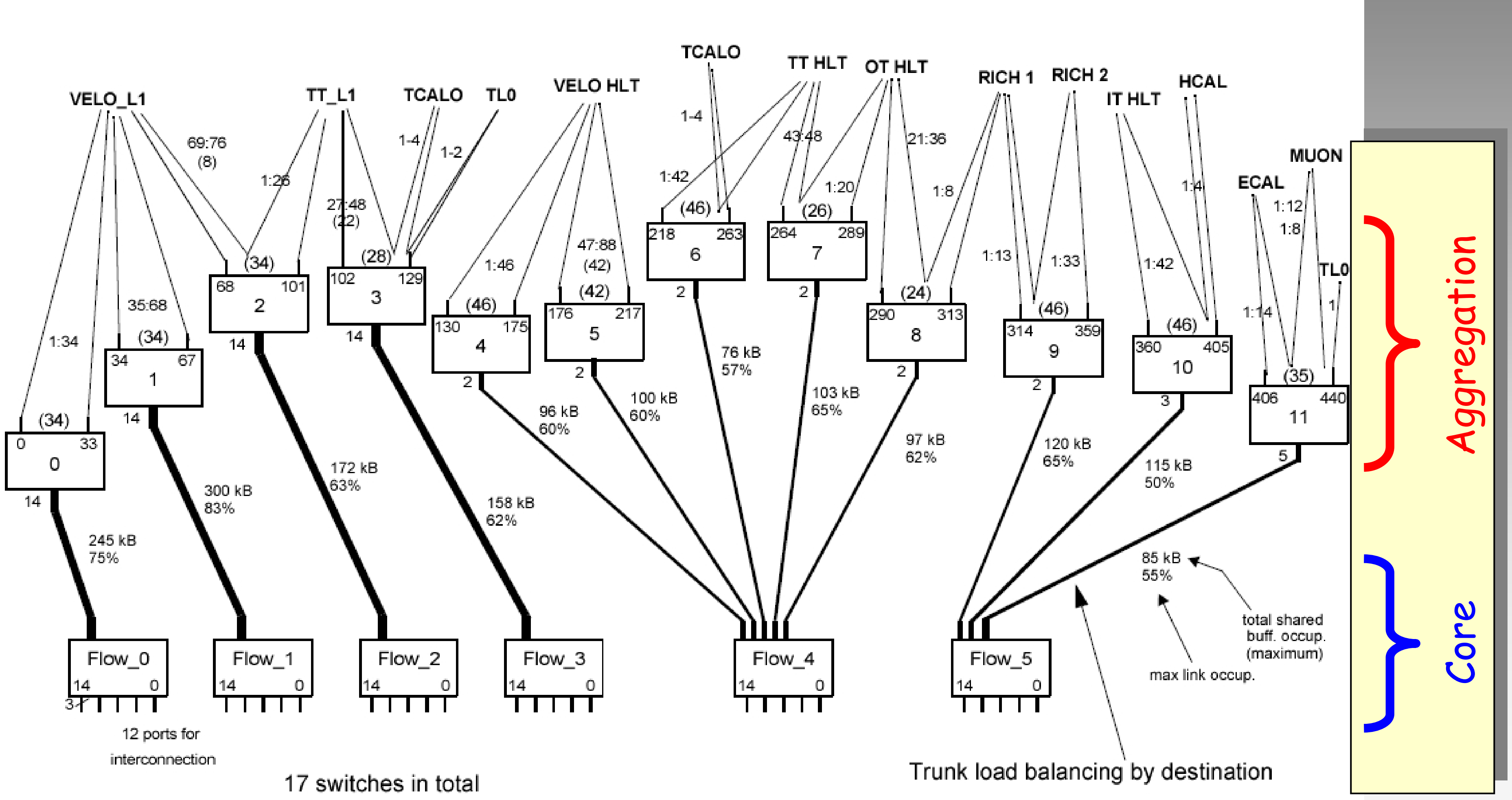


Full scale extrapolation: simulation

- Extrapolation to full scale system
  - Discrete time simulation
  - In-house development, C
  - MC produced data samples used as input, gives realistic
    - Frame timing
    - Frame sizes
- Started with generic switch model, interconnected with 1Gbps links
- Later refined to include
  - Priority queues
  - Link aggregation (link load balancing)
  - Internal switch architecture
  - Higher bandwidth interconnection (stacking) on internal links

Simulation model

Layout based on generic 48 ports switches



total shared buff. occup. (maximum)

max link occup.

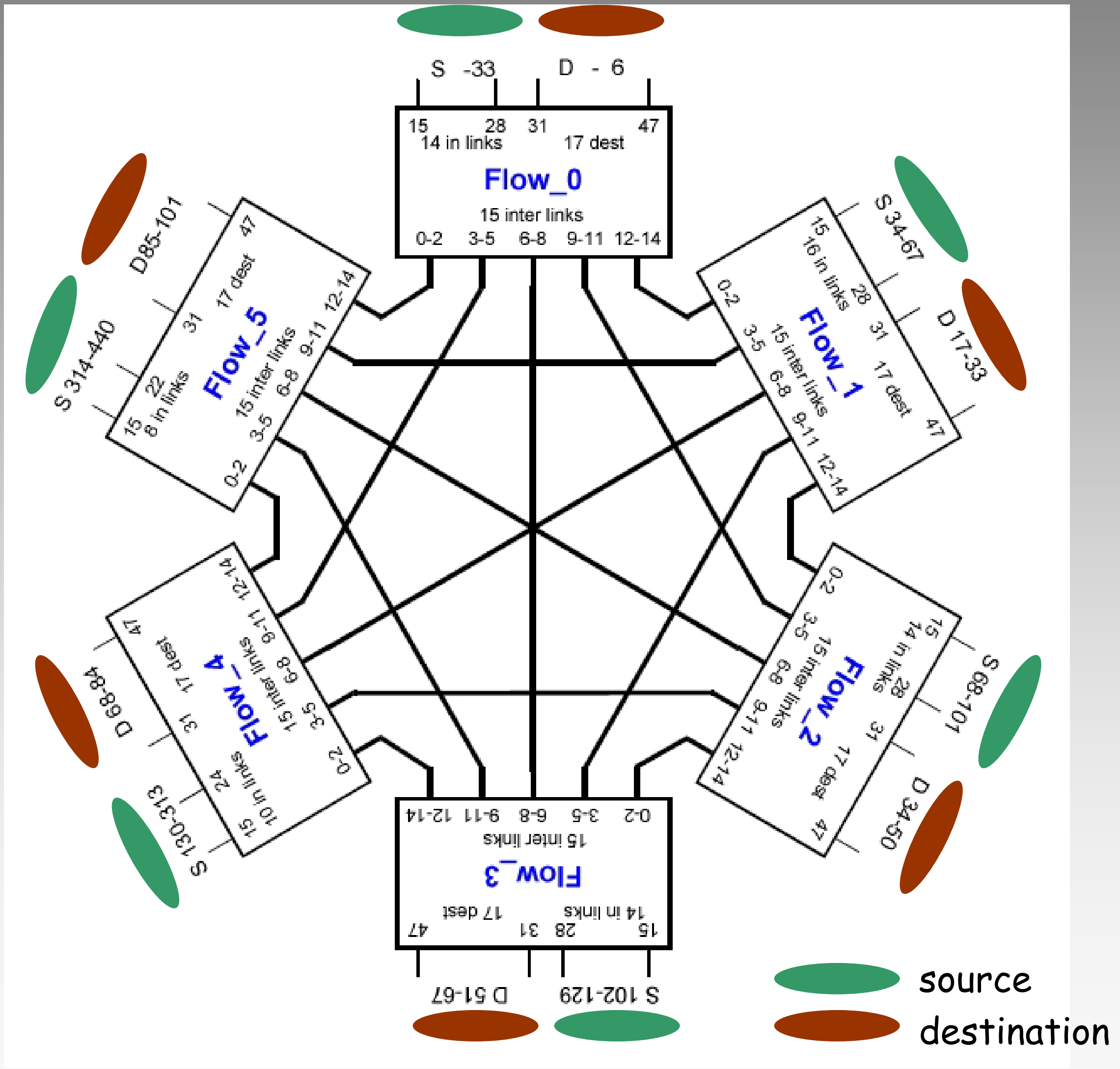
12 ports for interconnection

17 switches in total

Trunk load balancing by destination

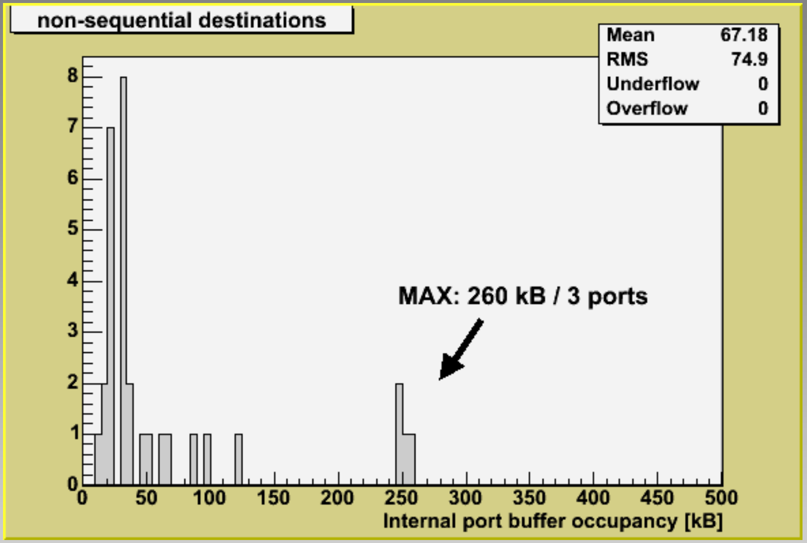
Simulation

- Generic switch model:
  - 48 ports
  - no speed-up in the fabric (96 Gbps fabric capacity)
- Internal connections:
  - Aggregated links with 3 GbE connections
  - Used in full-duplex
- Optimized destination port assignment improves memory utilization:
  - Force "next destination" to be on a different switch
- Single GbE connection to destination host
- Two independent flows for L1 and HLT traffic
- No priority queuing

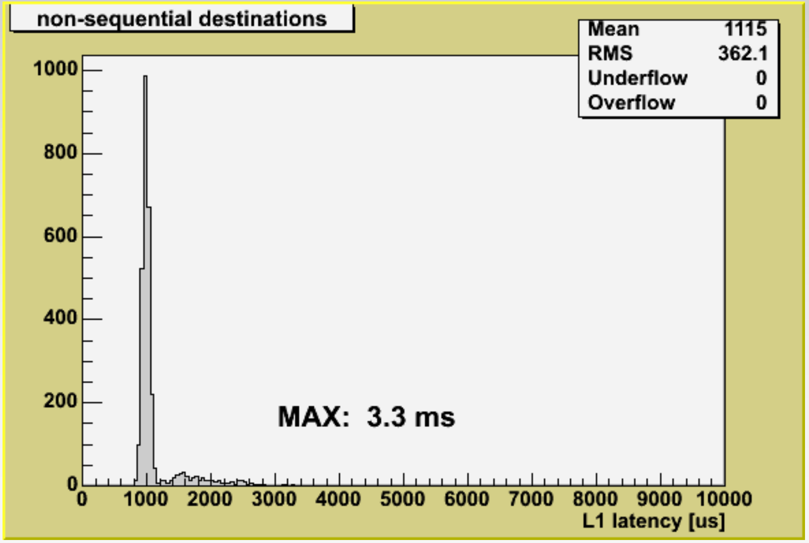


First simulation results

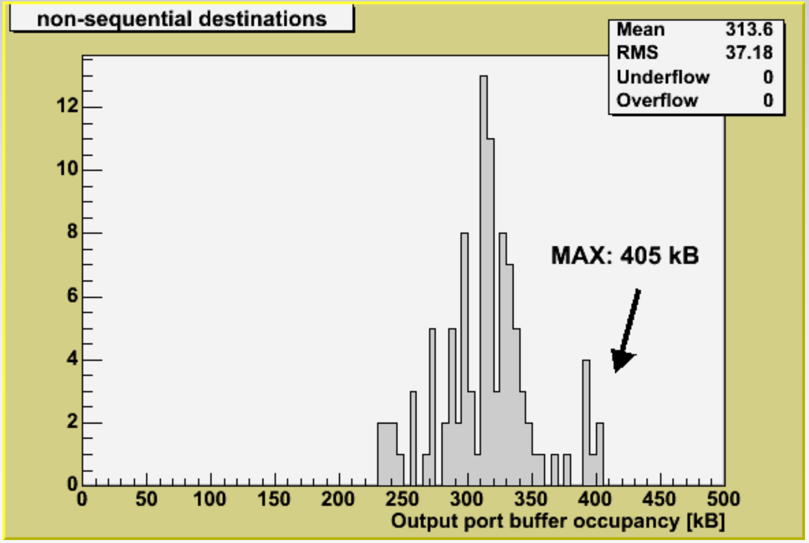
- The three most interesting values:
  - L1 event latency: < 4 ms
  - Internal buffer occupancy: < 260 kB / 3 ports
  - Output port buffer occupancy: < 405 kB / port



MAX: 260 kB / 3 ports



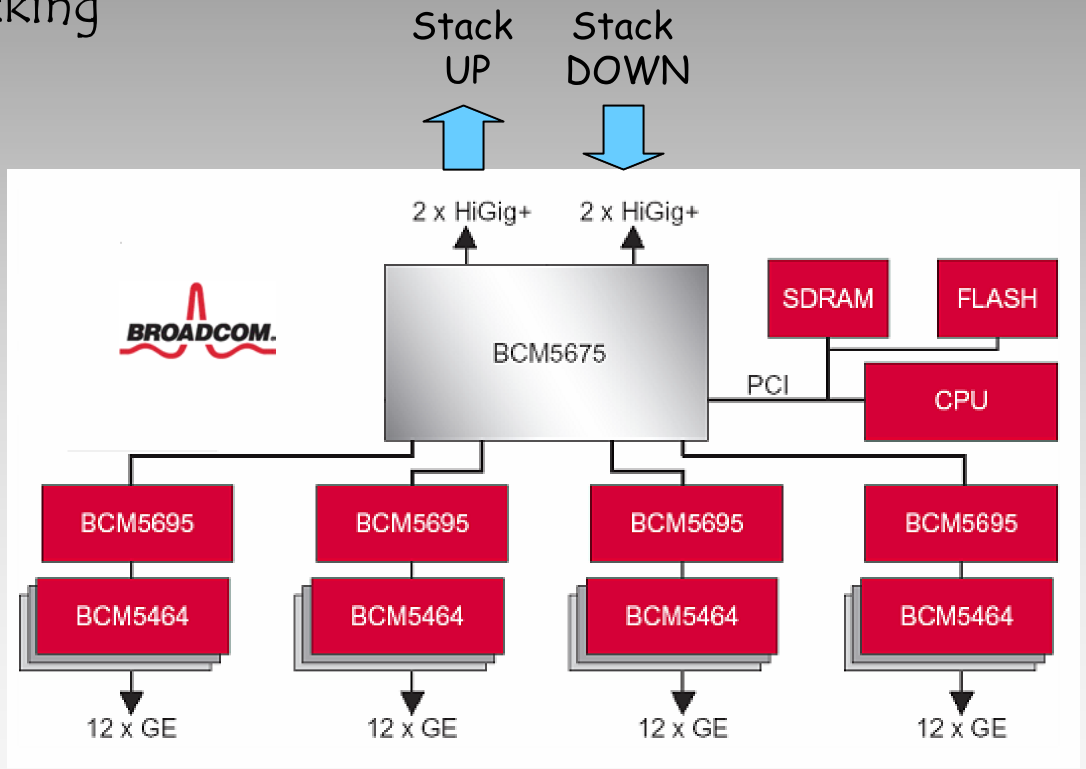
MAX: 3.3 ms



MAX: 405 kB

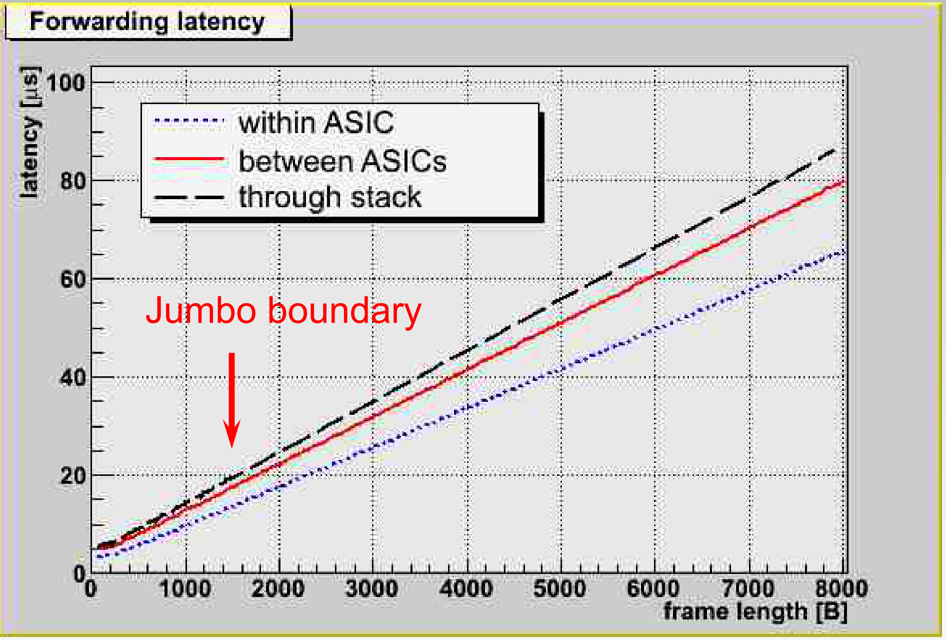
Model specific simulation

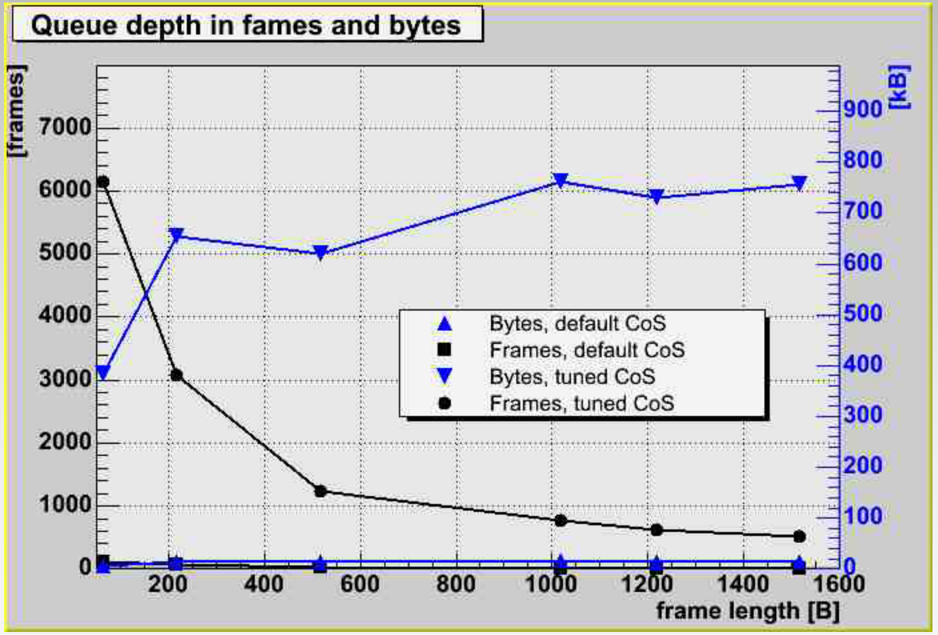
- Refined simulation to reflect the architecture of switch based on the Broadcom BCM5675/5695 chipset
- 48 GbE ports
- 2 x 20 Gbps stacking



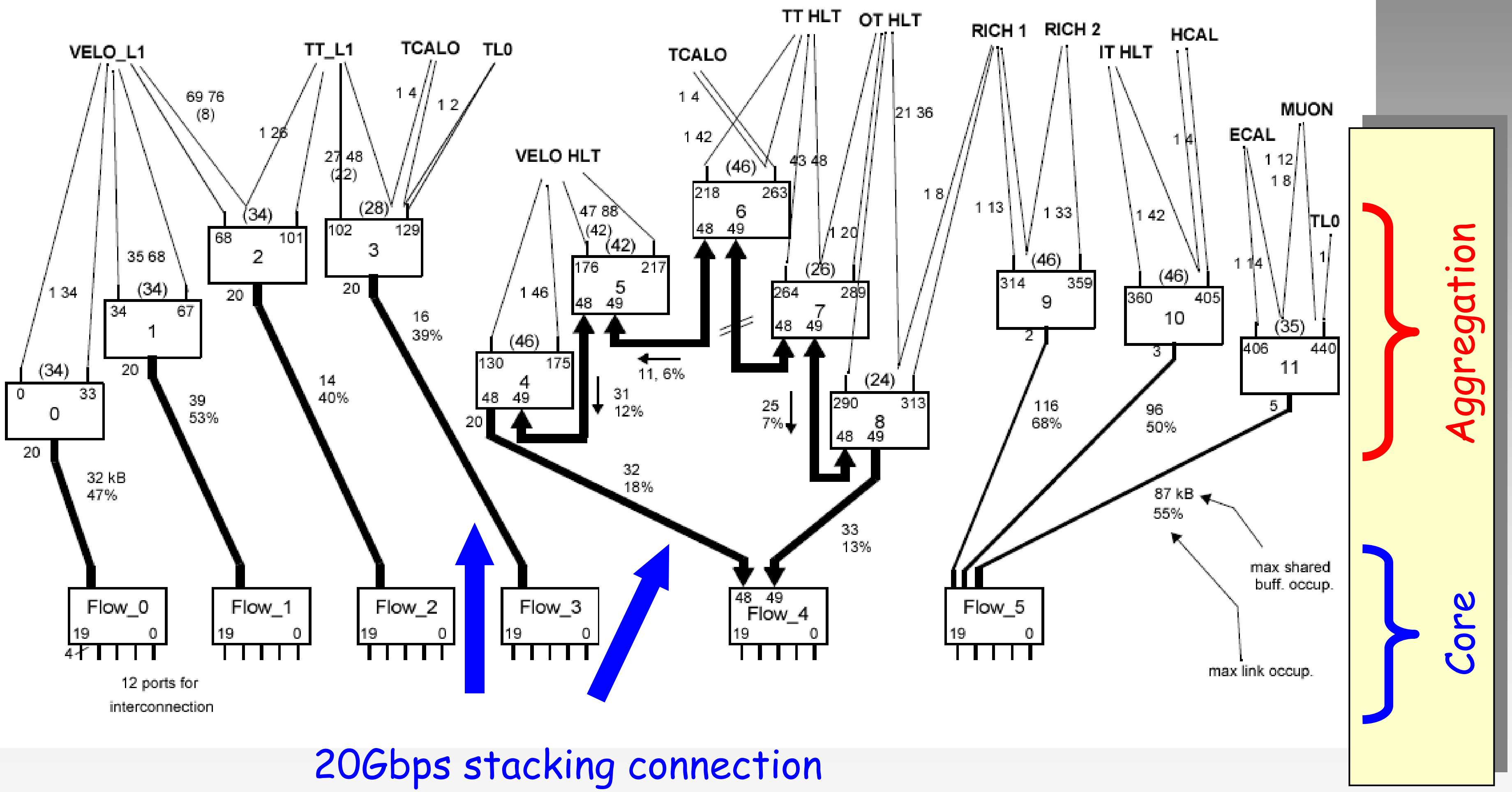
Known behaviour

- We have evaluated switches based on this architecture in our test-bed
  - Latency
  - Queue depths with different Class of Service settings
- Interesting feature: stacking for connection between aggregation and core layer





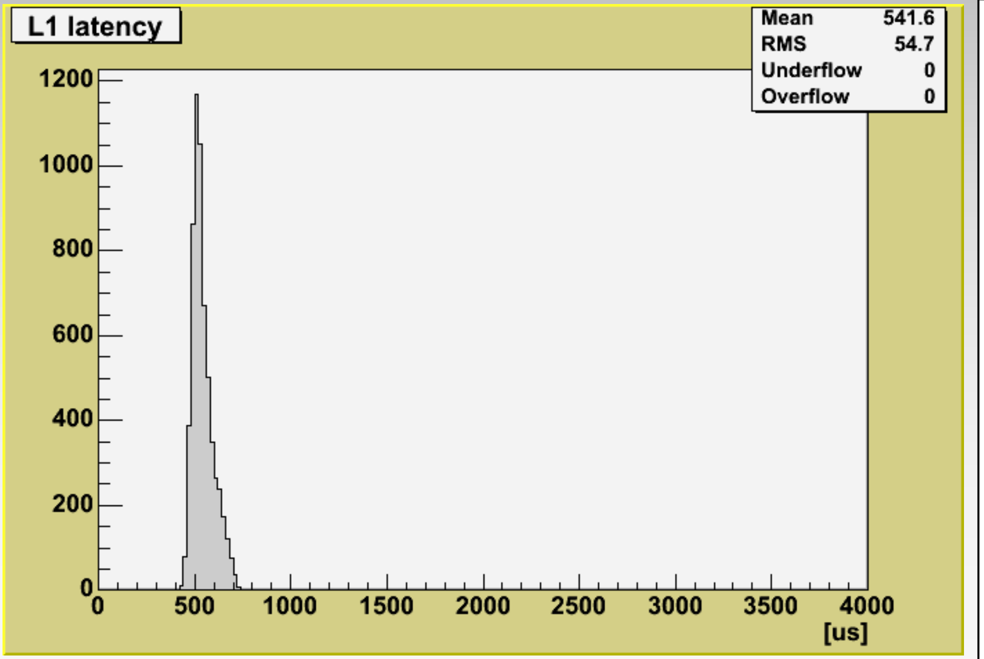
Refined simulation model

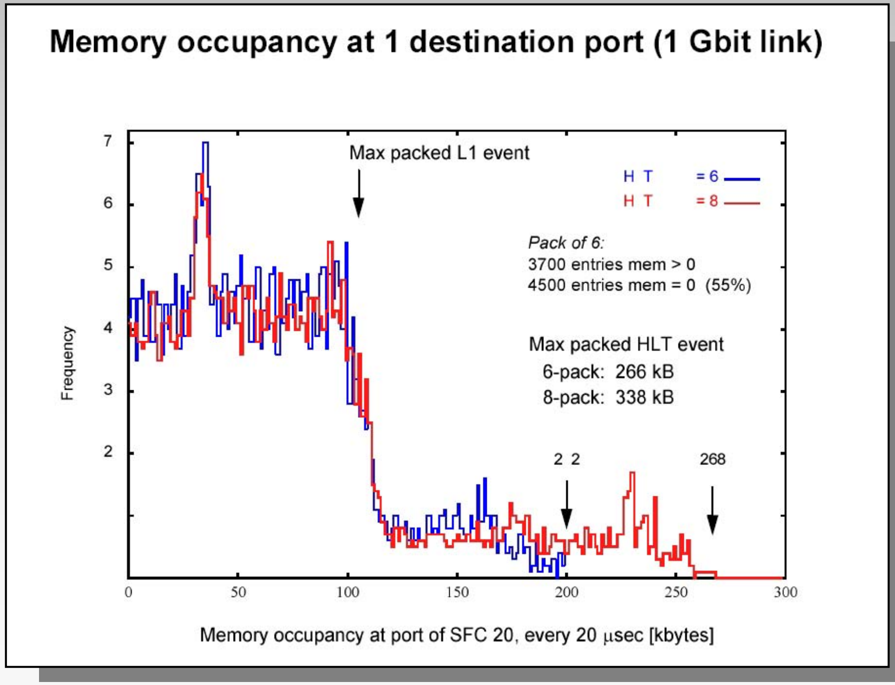


20Gbps stacking connection

Refined simulation results

- Additional changes:
  - 2 x 1 GbE links to destination (SFC)
  - 4 GbE in (internal) aggregated links
  - Two priority queues (L1 traffic prioritized over HLT)
- Outcome:
  - Lower L1 latency: < 1 ms
    - Due to increased bandwidth on all connections (stacking, internal and to destination)
  - While keeping memory utilization low on output ports: < 400 kB
    - Within the limits of available memory



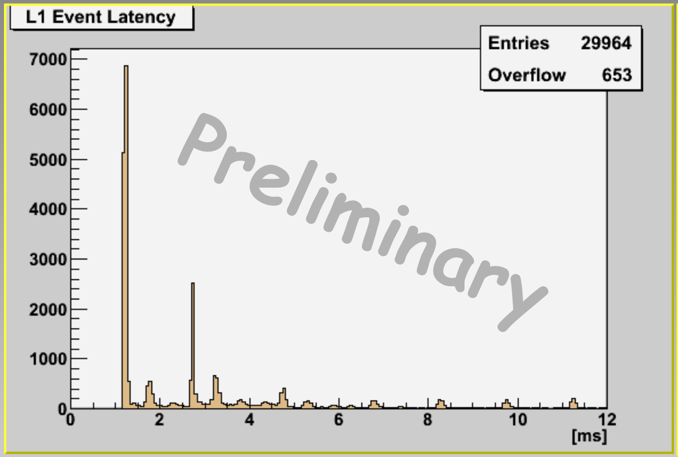


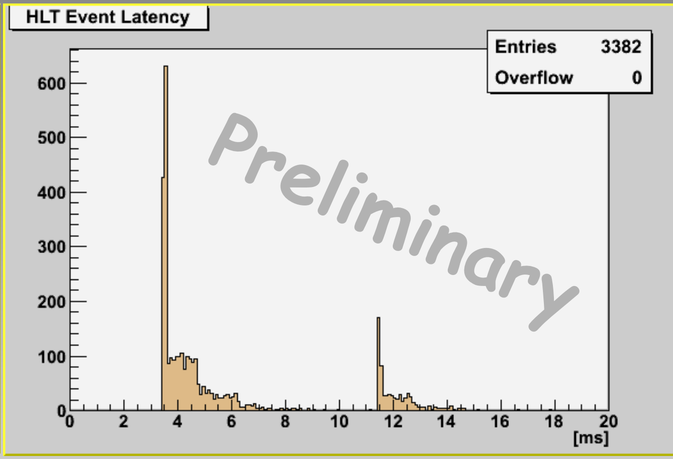
Single switch simulation

- The arrival of large port density switches on the market raised our interest in the single switch solution
- Important requirements:
  - Non-blocking
  - Over-commitment factor < 2
    (Note that LHCb DAQ traffic is uni-directional!)
- A preliminary study indicates this type of switch can be used, and available on the market
- Devised a simulation model based on an existing switch
  - Cross-bar fabric
  - Up to 96 GbE ports per blade (→ over 1200 ports in total)
  - 128 MB buffer memory per blades
  - LHCb timing for L1 and HLT traffic
  - Overlaid 20% large events
  - Single GbE link to destination
- Studied two cases:
  - No priority queues
  - Two priorities: high priority for L1 traffic

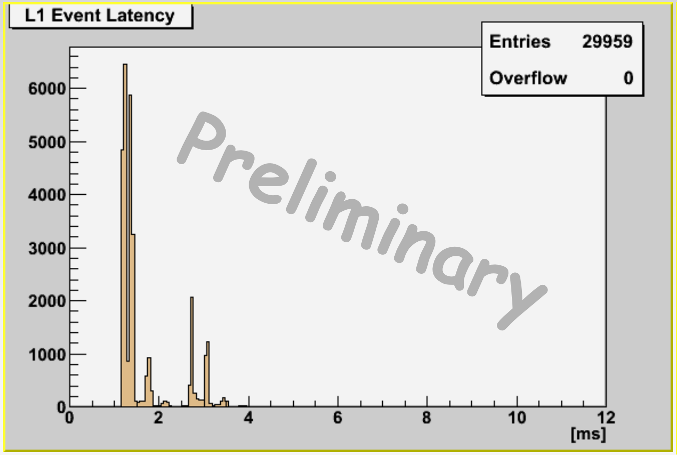
Preliminary simulation results

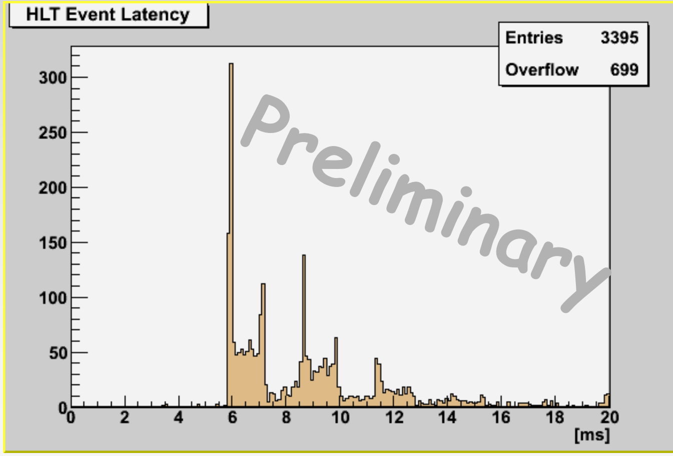
- No priority classes:
  - Memory utilization ~7 MB / blade (128 MB available)
  - L1 traffic can be queued behind HLT traffic





- Two priority queues:
  - Reduces L1 latency below 5 ms (below 2 ms for normal events)
  - Memory utilization raises insignificantly to ~8 MB / blade





Summary

- The LHCb DAQ test bed has been used to evaluate Gigabit Ethernet switch performance
  - Foundry, Nortel, Force10, Extreme, Cisco, etc…
- Typical performance figures
  - Forwarding latency
    - Edge: 15-20 μs (1500B), ~60 μs (9000B)
    - Core: ~50 μs (1500B), ~100 μs (9000B)
  - Loss rates under LHCb traffic pattern are below 10^{-10} frames for good candidates
  - Typical queue depths (frame size dependent)
    - Edge: ~100 kB
    - Core: up to ~4 MB
  - Quality of Service settings in some switches allow to use larger portions of SHARED memory
    - Up to ~800 kB per port in edge switches
    - Up to ~4MB per port in core switches
- Feedback from test-bed was used to refine our simulation model used to predict the performance of the full-size setup
- Simulation models give us predictions of
  - Level 1 event latency well below 10 ms (below 1 ms in extreme case)
  - Memory requirements below 400 kB per egress queue
- The needs of the LHCb readout network are met by high performance GbE switches with the available features (quality of service, link aggregation, stacking)