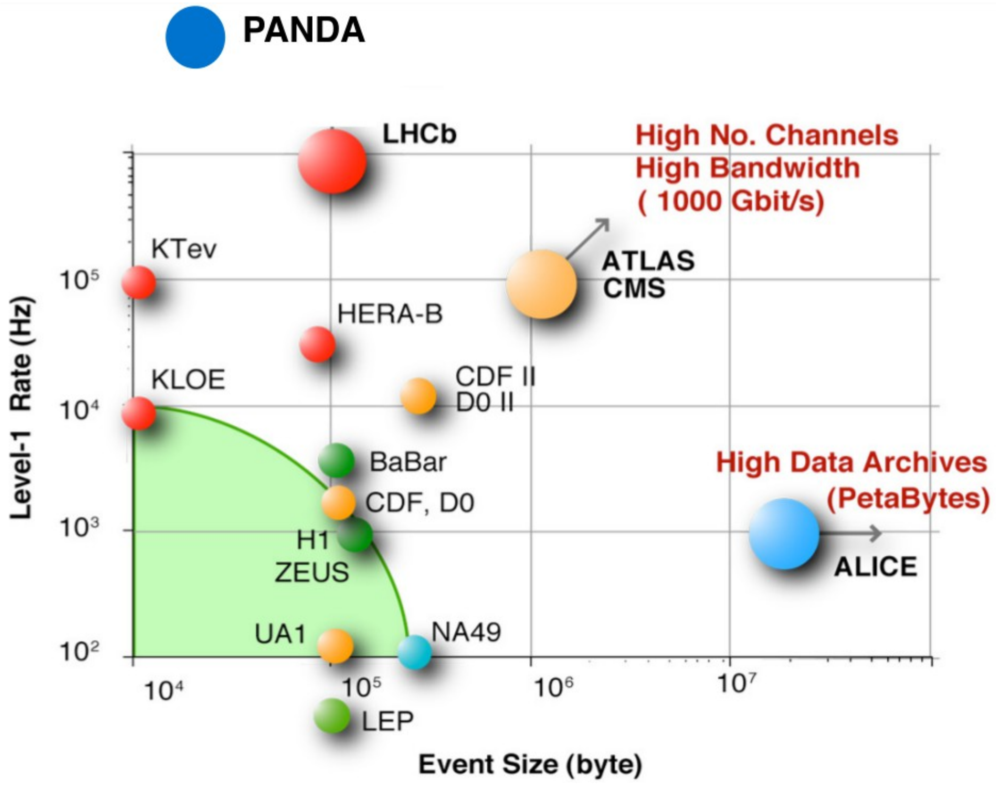
Preliminary results of PANDA DAQ System Proposal

Mateusz Michałek
Cracow University of Technology,
Institute of Nuclear Physics PAN

PANDA outline

- 200GB/s of average throughput (initially)
- 20e6 interactions per second
- 400 detector FrontEnds (initially)
- Unknown number of event building and filtering farms
- FrontEnd electronics monitors signals from detectors and in case of crossing a threshold datapacket is formed and sent to concentrator which then forwards aggregated info to event building node
- There is no hardware triggering. All the data is processed and usefulness of event is estimated after event building and filtering run on full event data
- DAQ system should provide way to deliver all fragments of the same farm

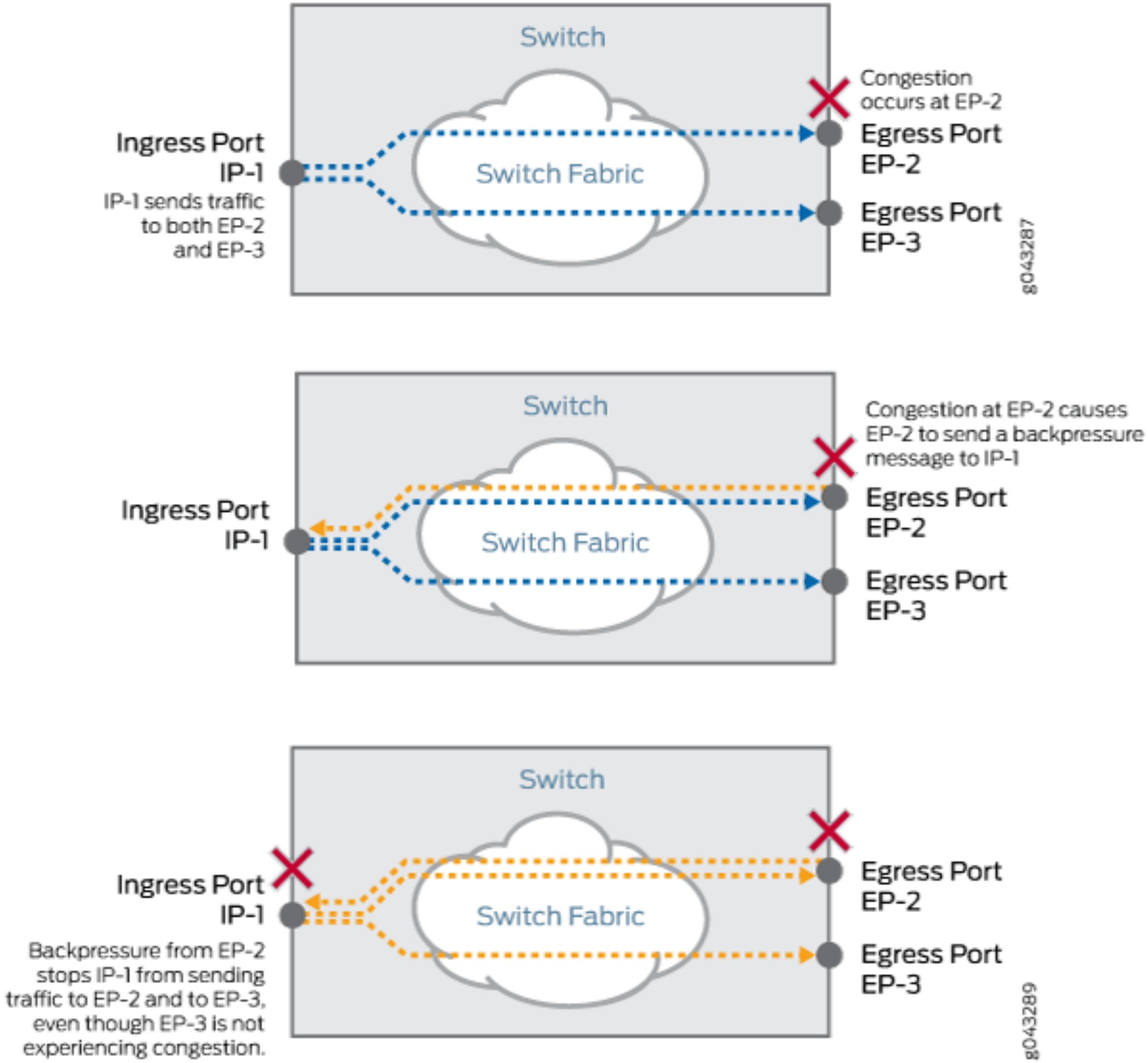


Classical congestion problem

Typical network architecture avoids packet dropping by creating backpressure on ingress port. This approach is based on assumption that traffic is distributed evenly over the network and sender has large enough buffer.

PANDA DAQ role is to deliver all fragments of data generated by detectors during one epoch (duration of 2 us) to one farm.
Therefore traffic shape is vastly different and there are high spikes of data on egress port instead.

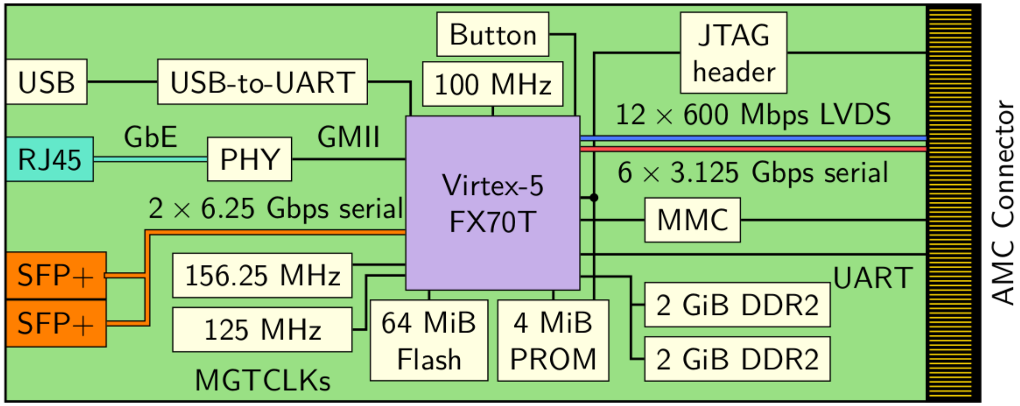
Network implemented in this typical way is not suitable for PANDA DAQ because congestion on single egress port will corrupt data of following events, even though these events are not sent to congested port.



Switch
Switch Fabric
Ingress Port IP-1
IP-1 sends traffic to both EP-2 and EP-3
Congestion occurs at EP-2
Egress Port EP-2
Egress Port EP-3
g043287
Switch
Switch Fabric
Ingress Port IP-1
Congestion at EP-2 causes EP-2 to send a backpressure message to IP-1
Egress Port EP-2
Egress Port EP-3
Switch
Switch Fabric
Ingress Port IP-1
Backpressure from EP-2 stops IP-1 from sending traffic to EP-2 and to EP-3, even though EP-3 is not experiencing congestion.
Egress Port EP-2
Egress Port EP-3
g043289

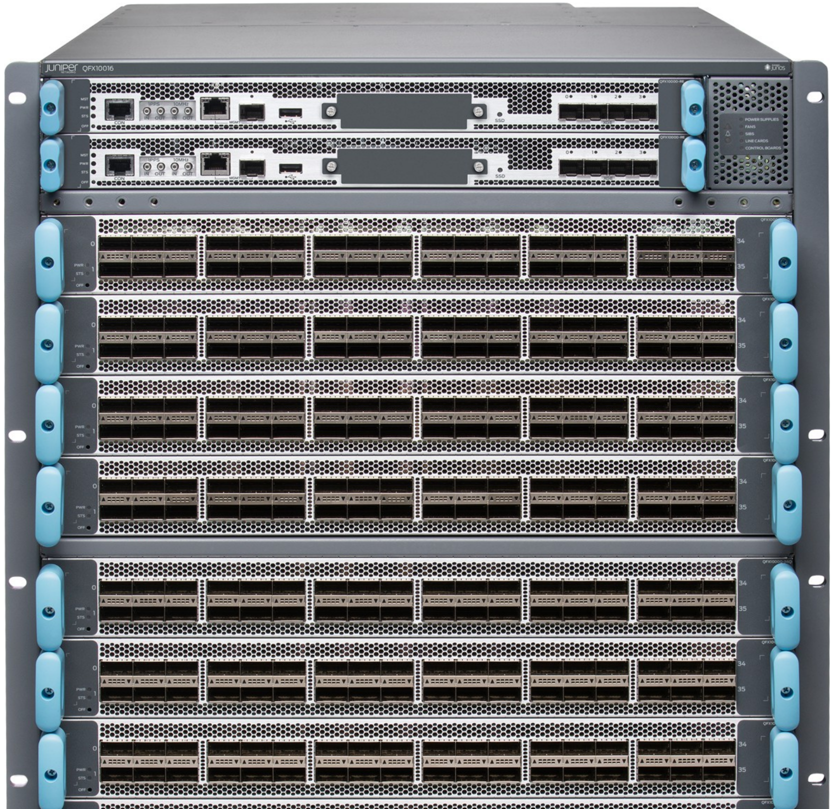
ATCA approach

4 AMC cards with 2 6.25Gbps SFP+ ports are fitted into carrier boards.
13 carrier boards are fitted into ATCA crate.
Each ATCA crate provides 104 external links.
8 crates needed to connect FrontEnds and farms, not counting crate-crate interconnects.
FPGA's allows to address congestion problem.
Scaling needs rearranging interconnects.

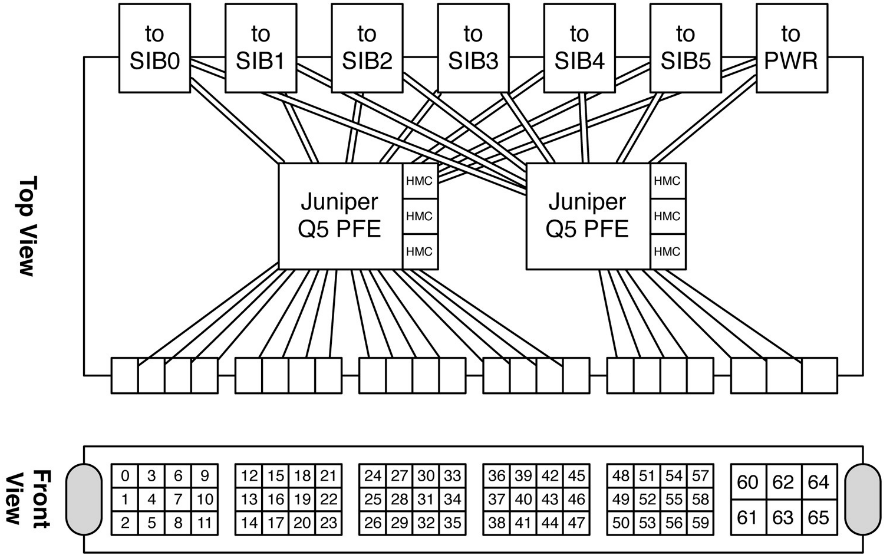


Approach using off-the-shelf hardware (Juniper QFX10016 ethernet switch)

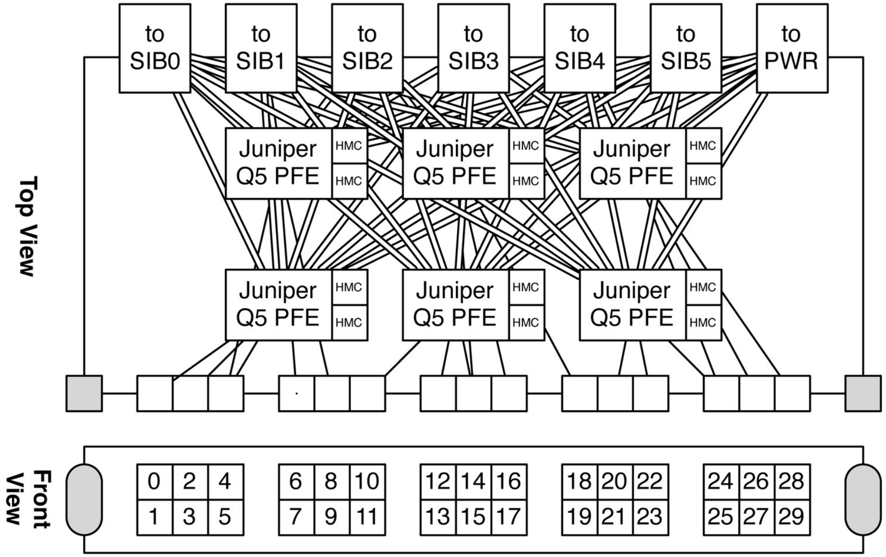
QFX10016 allows fitting 16 line cards.
Two example cards shown on the right.
QFX10000-60S-6Q card provides 60 10Gbps links and 6 40Gbps links.
QFX10000-30C card provides 30 100Gbps links
Each Q5 chip is tightly coupled with 4GB of packet memory



Juniper QFX10000-60S-6Q Line Card

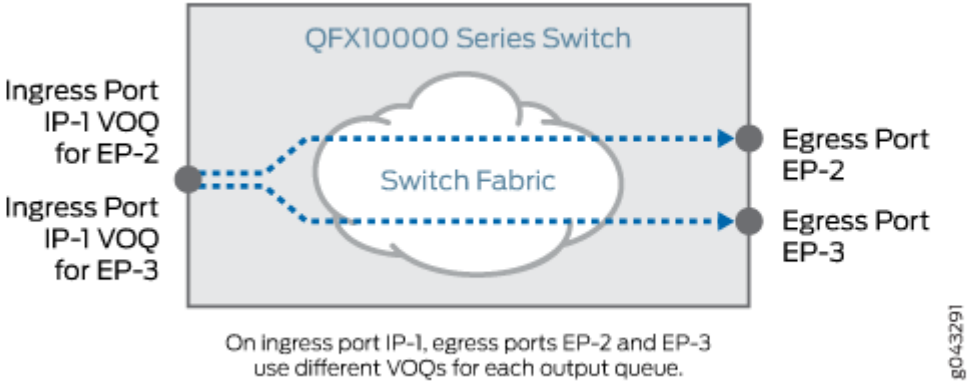


Juniper QFX10000-30C Line Card

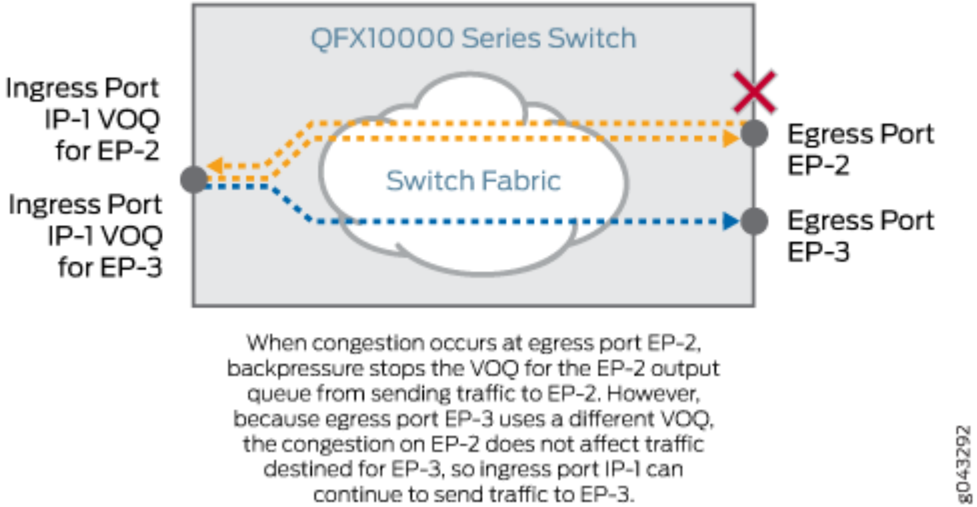


Congestion problem using Virtual Output Queue

Juniper QFX10k switches are using VOQ to distinguish congested egress port at ingress port. This technique allows to create backpressure on sender regarding destination address of incoming data.



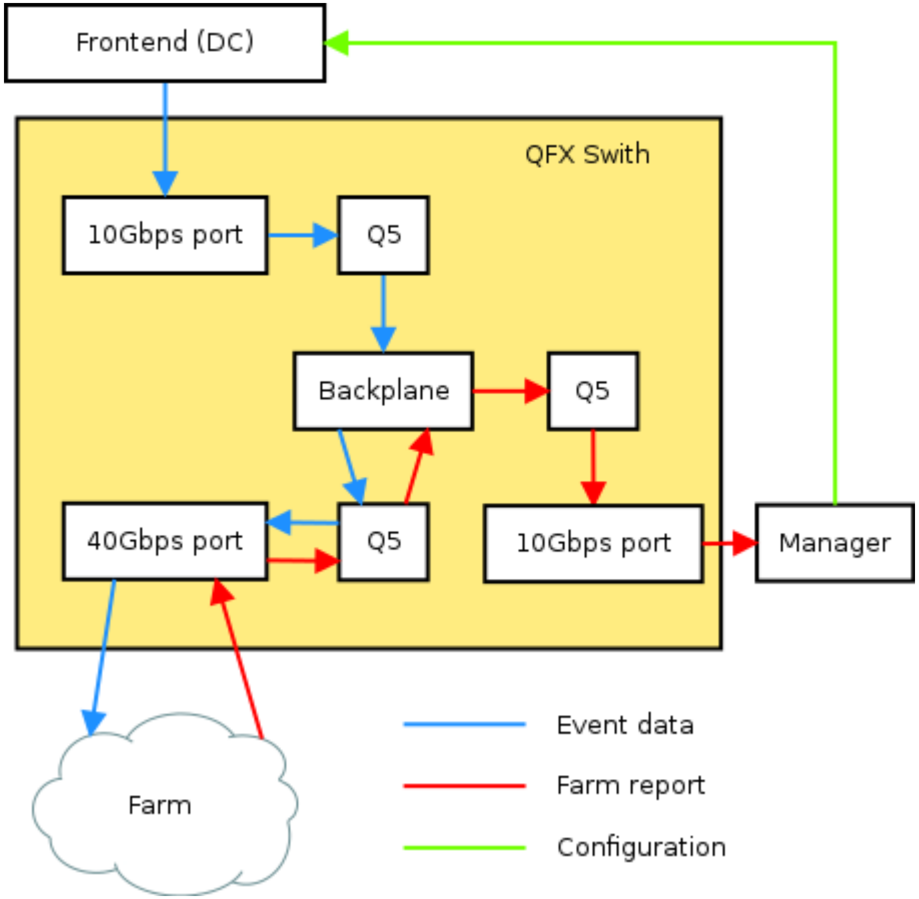
On ingress port IP-1, egress ports EP-2 and EP-3 use different VOQs for each output queue.



When congestion occurs at egress port EP-2, backpressure stops the VOQ for the EP-2 output queue from sending traffic to EP-2. However, because egress port EP-3 uses a different VOQ, the congestion on EP-2 does not affect traffic destined for EP-3, so ingress port IP-1 can continue to send traffic to EP-3.

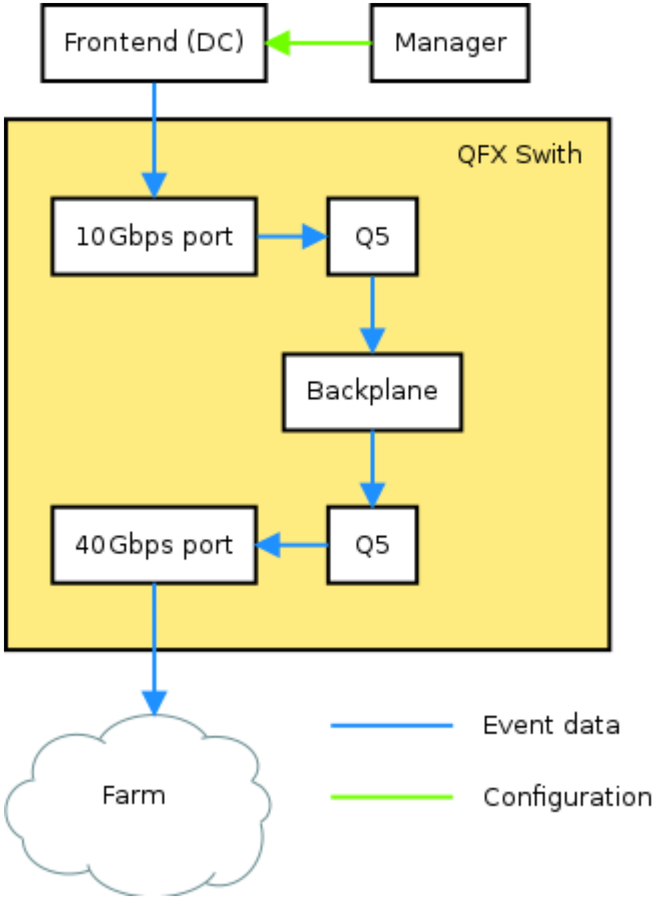
“Farm queue balancing” addressing scheme

- •Concentrator modules are connected via 10Gbps ports.
- •Event building farms are connected to 40 Gbps ports.
- •There is one manager module which is connected to 10Gbps port
- •All FrontEnds are synchronized by Manager module.
- •Farms put received data fragments in buffer and after event data is complete, event is inserted into queue.
- •Farms are sending reports to manager on every change in event-building queue.
- •FrontEnds are sending event data to farm according to address commanded by manager module.
- •Manager selects destination address basing on queues.
- •Destination scattering is enabled to avoid egress port congestion in case event-building times are negligible in contrast to transmission time.



“Round-Robin” addressing scheme

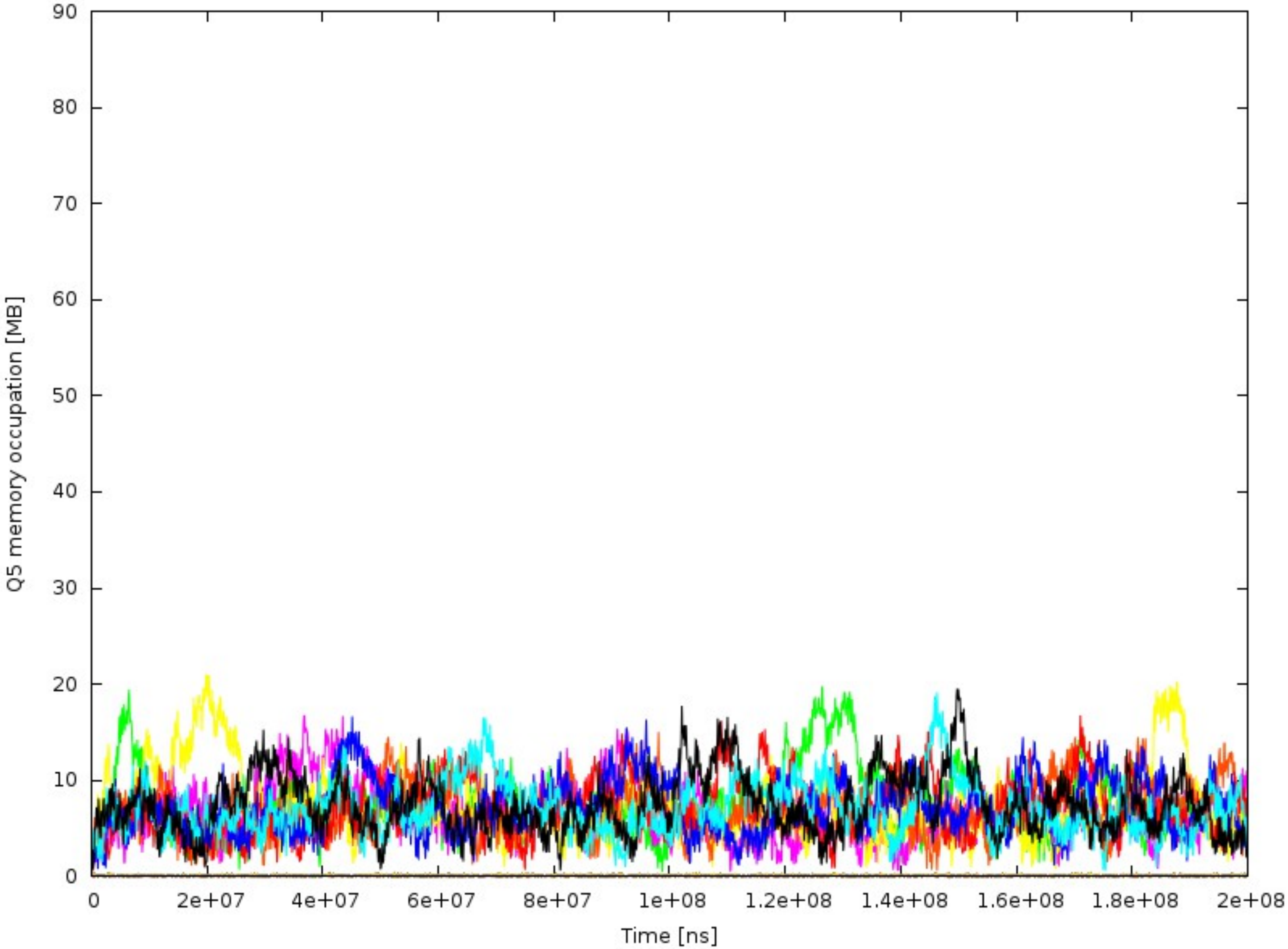
- Concentrator modules are connected via 10Gbps ports.
- Event building farms are connected to 40 Gbps ports.
- There is one manager module which is not connected to switch
- All FrontEnds are synchronized by Manager module.
- Farms put received data fragments in buffer and after event data is complete, event is inserted into event-building queue.
- FrontEnds are sending event data to farm according to address commanded by manager module.
- Manager selects destination according to Round-Robin.



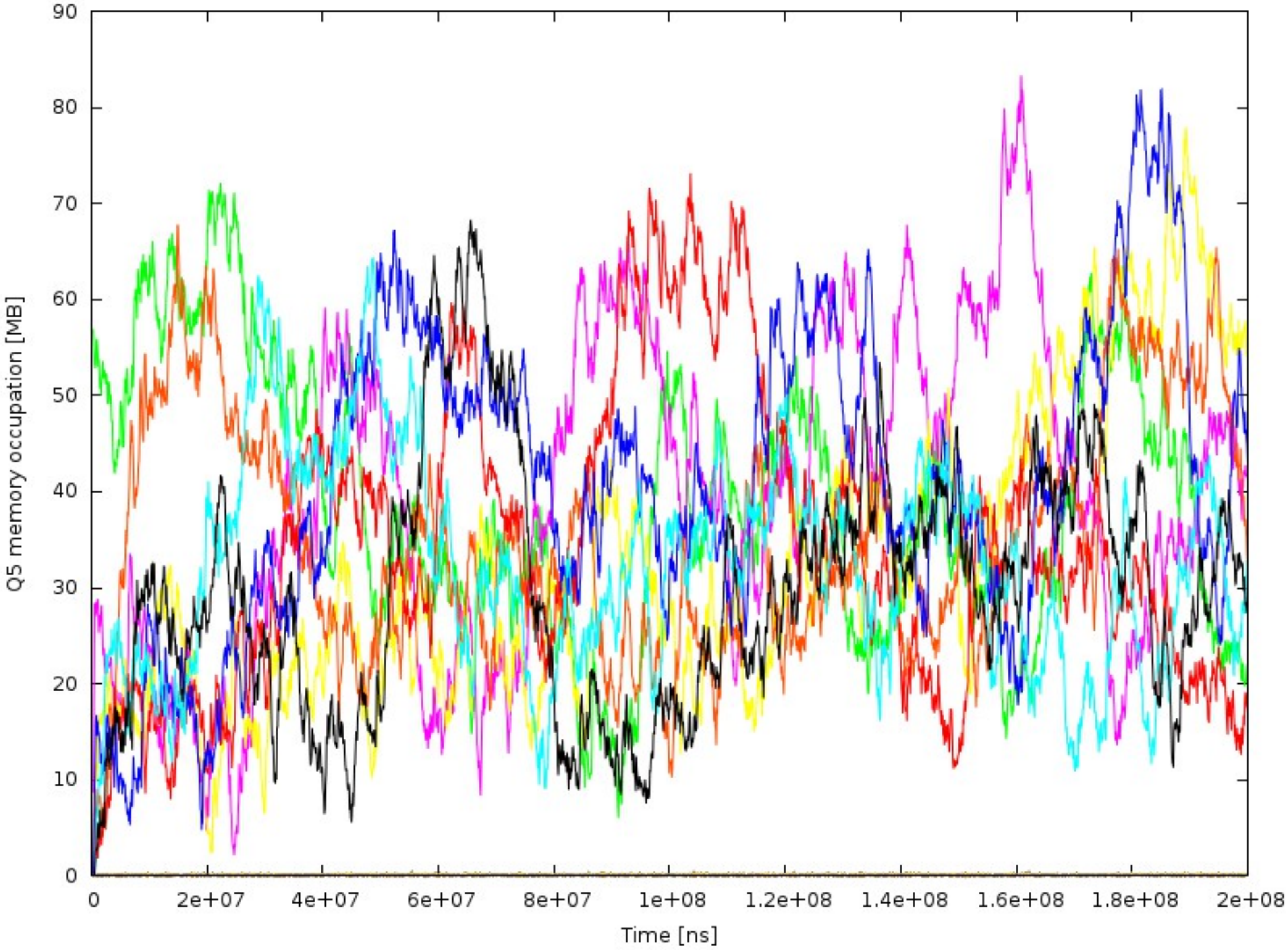
Simulation

- Gaussian distribution of event data length and event-building time
- 1100 bytes per packet
- 2400ns between packets
- 223GBps throughput
- 500 FrontEnds
- 50 farms
- Both addressing schemes tested

Q5 memory occupation versus time for "Round-Robin" addressing scheme



Q5 memory occupation versus time for “queue balancing” addressing scheme



Conclusions

- Total throughput is sufficient
- Easy expansion and roughly 40% margin (9 out of 16 slots populated with 60S-6Q cards)
- Round-Robin addressing gives equalized memory utilization but may lead to farm queue overload.
- In case of queue balancing scheme, Q5 Buffers are big enough to handle closed loop control delay caused by reporting.