Dataflow and Condition Data

4th ALICE ITS upgrade, MFT and O2 Asian Workshop 2014 @ Pusan

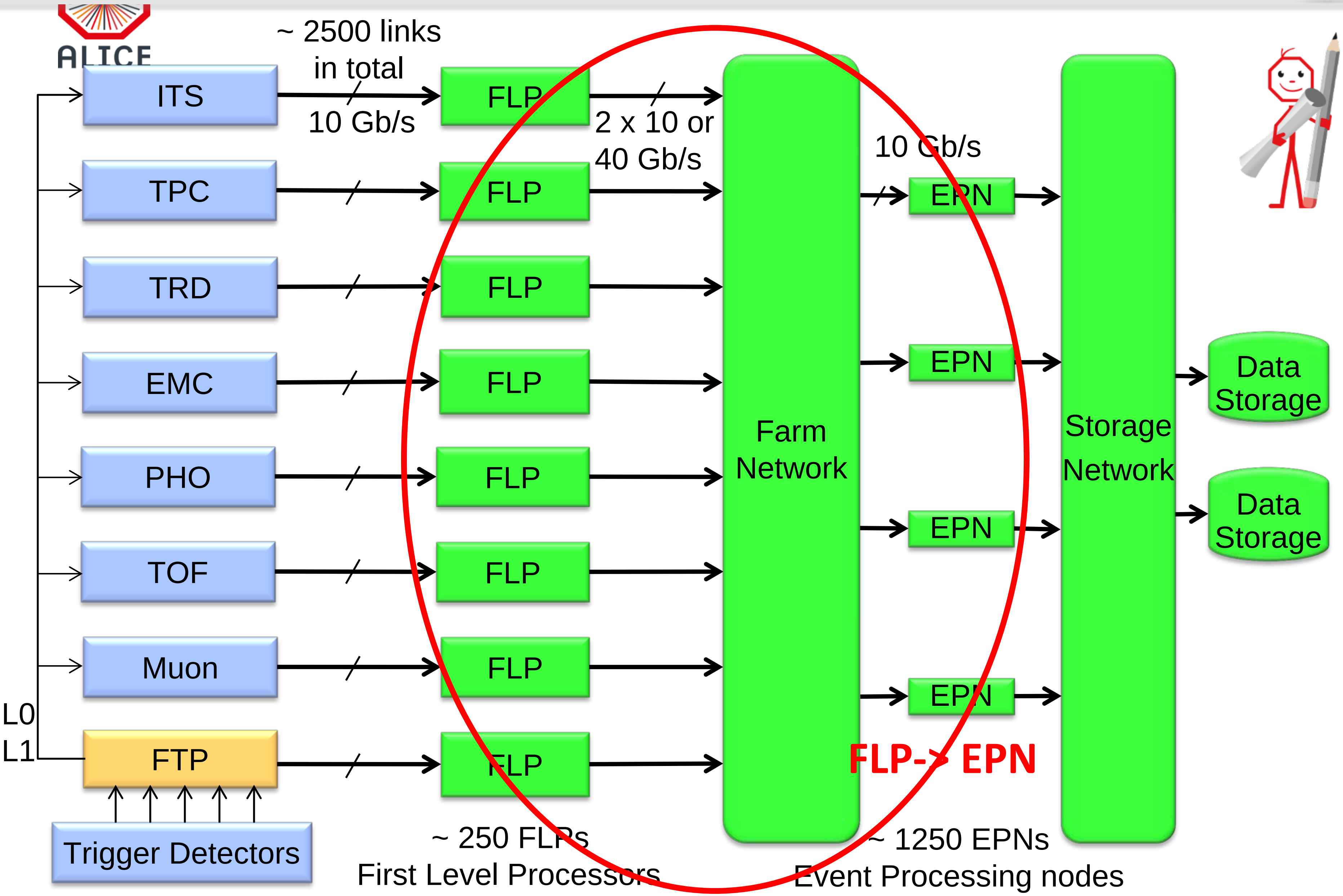
16 December 2014

Iosif Legrand

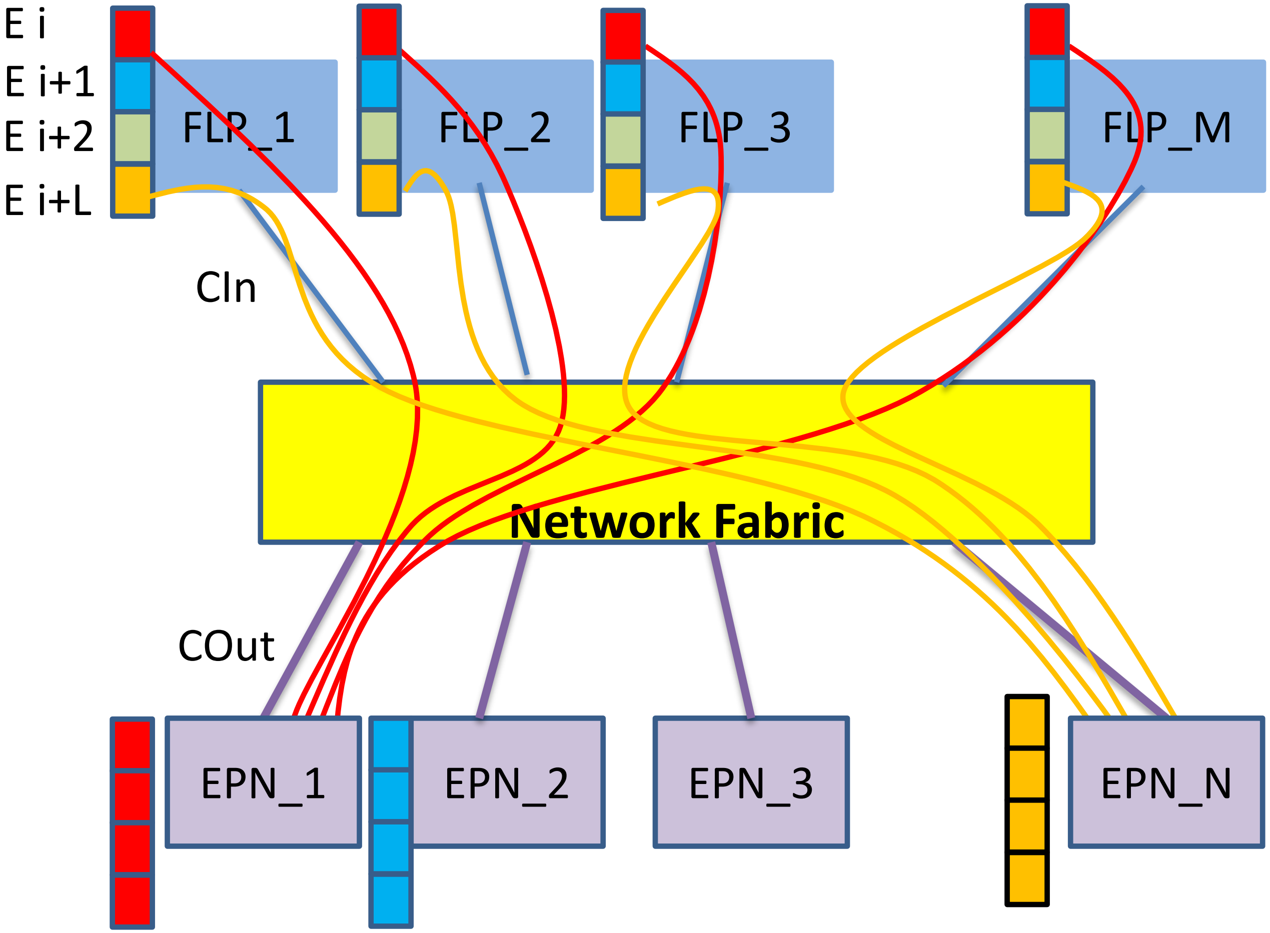
Outline

- Architecture considerations for the data flow
- Simulations and Modelling
- Cost estimations for different architectures
- Prototype system measurements
- Calibration data flows
- Summary

O² Hardware System



Traffic Pattern for FLP – EPN ; Time Frame Building



FAN IN PROCESS FOR EACH TIME FRAME

MUST BE DONE IN PARALLEL !

FLP Buffer Size

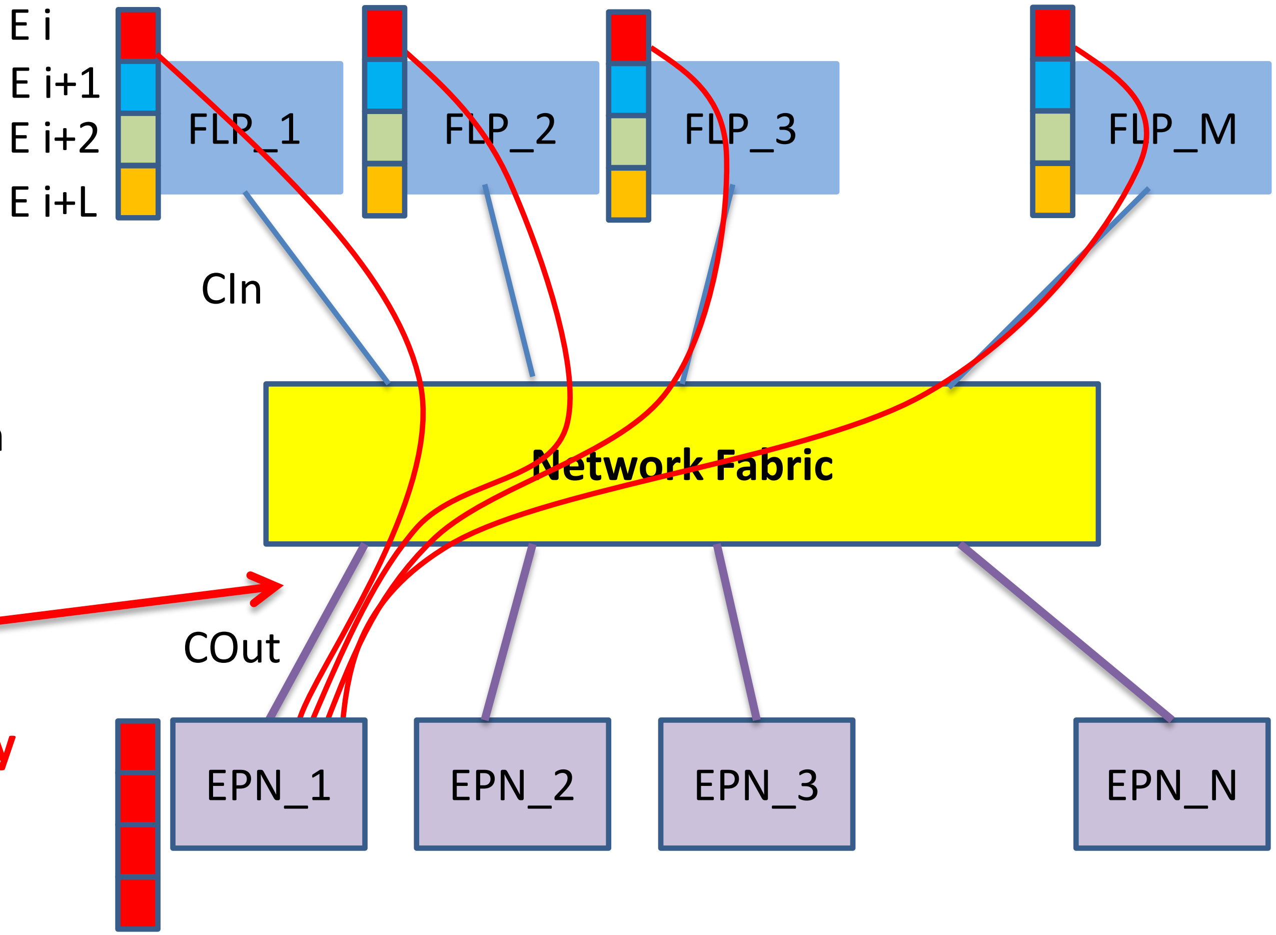
DpE = $\sum$ data_framgments

Latency~=DpE / Cout

Buffers ~= Event_rate *Latency*DP

Assuming all FLP send data in parallel to L EPNs (L >=M)

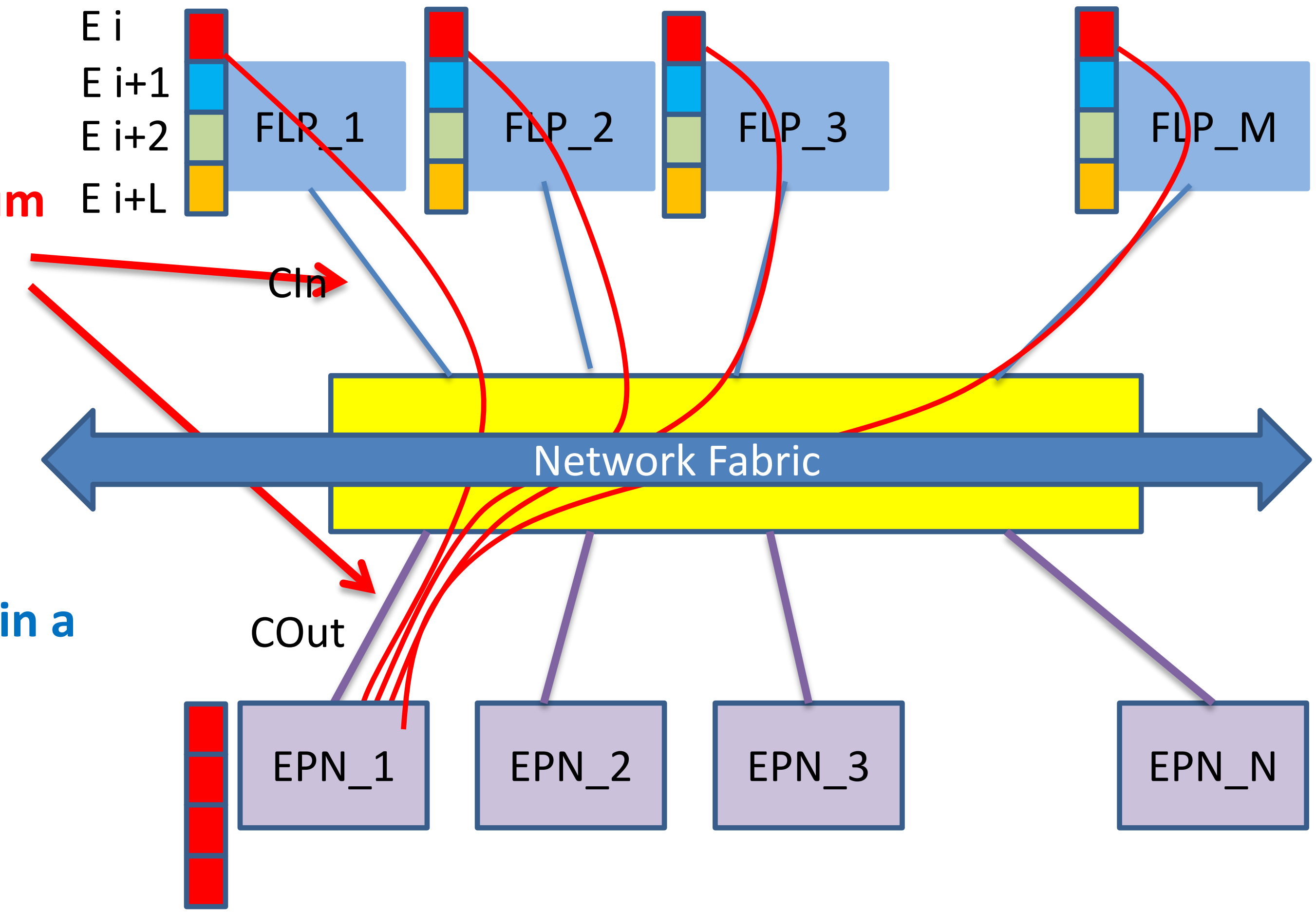
The capacity of the receiving links is the key element for the total latency and the amount of buffer size in FLP



The Number of Concurrent ”Event Building” processes

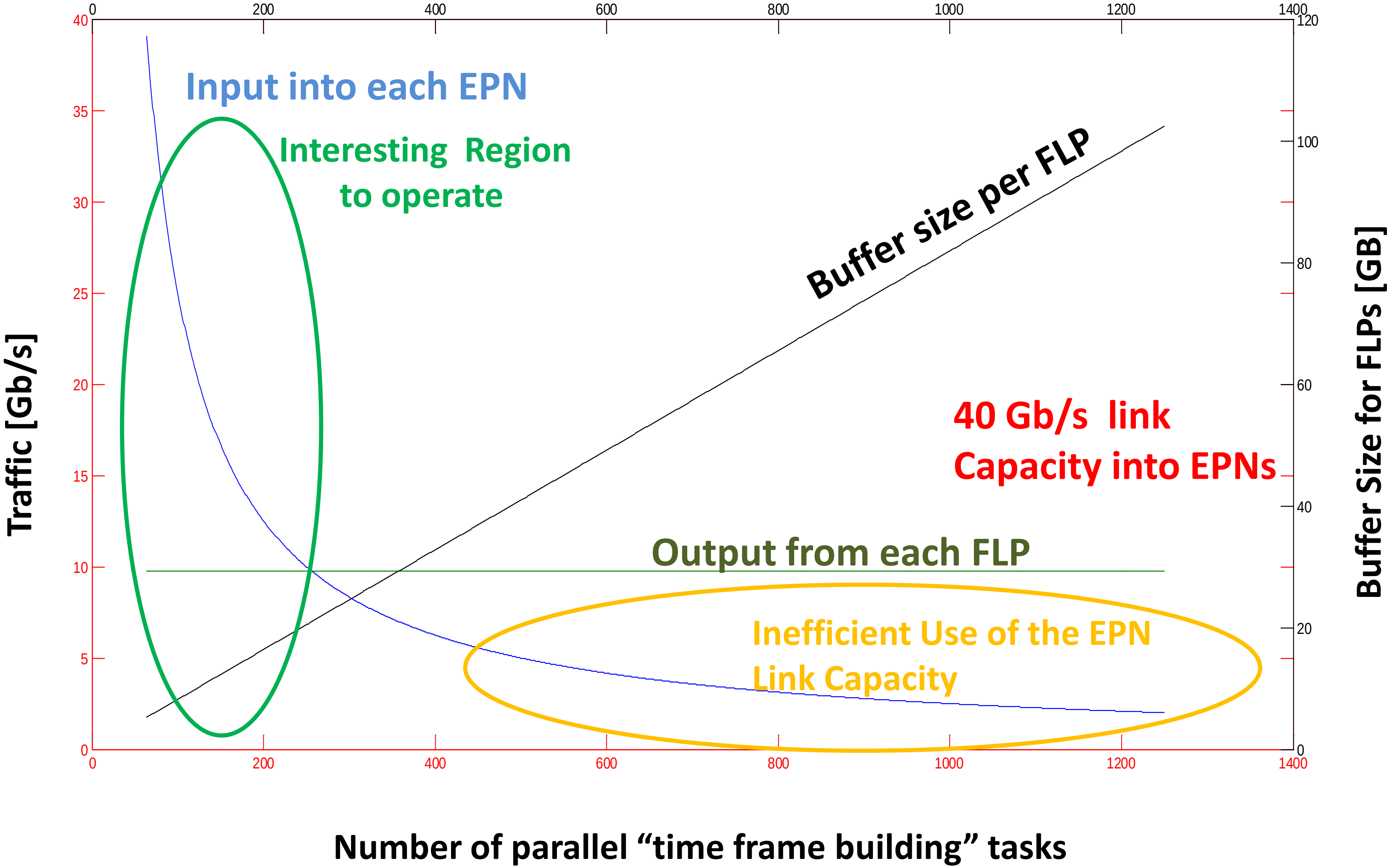
All FLP send data in parallel to L - EPNs

The capacity of the Sending and Receiving links define the minimum number of concurrent “Time Frame Building” tasks



**The bisection traffic should support the maximum throughput in a non blocking way for the entire system.
~ 2.5 Tb/s - 5 Tb/s**

Estimated average traffic per links and buffer capacity vs number of parallel transfers



Network Topology for Data Flow

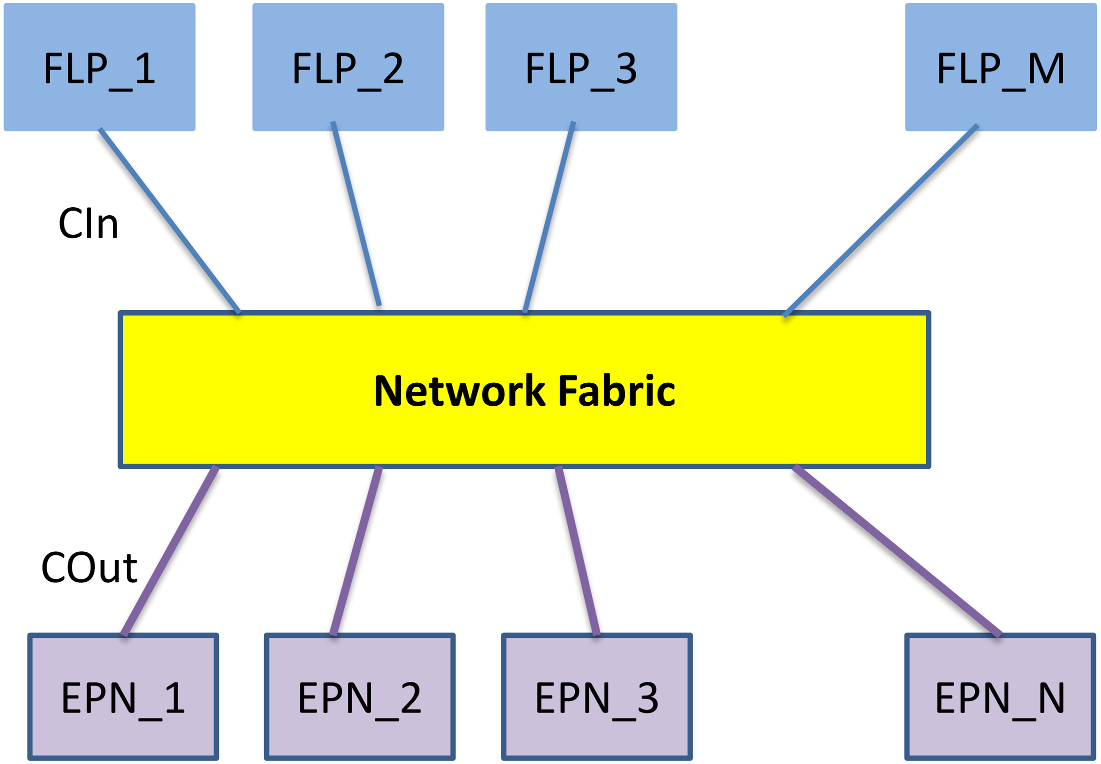
- ❖ High speed link capacity into EPN -> reducing the latency (memory buffers) and number of parallel transfers .

- ❖ Large number of EPNs into the "time frames building" switching fabric increase the cost of the switch and make the average traffic per EPN quite small.

- ✓ A two tier system, that does "time frame building" and than performs the EPN data Processing task should be considered, and it may provide a more cost effective solution.

- ✓ Need cost estimates for different switching technologies to evaluate different architectures . For each architecture we need too consider a set of possible algorithms to properly used to hardware design.

FLP – EPN Topology 1

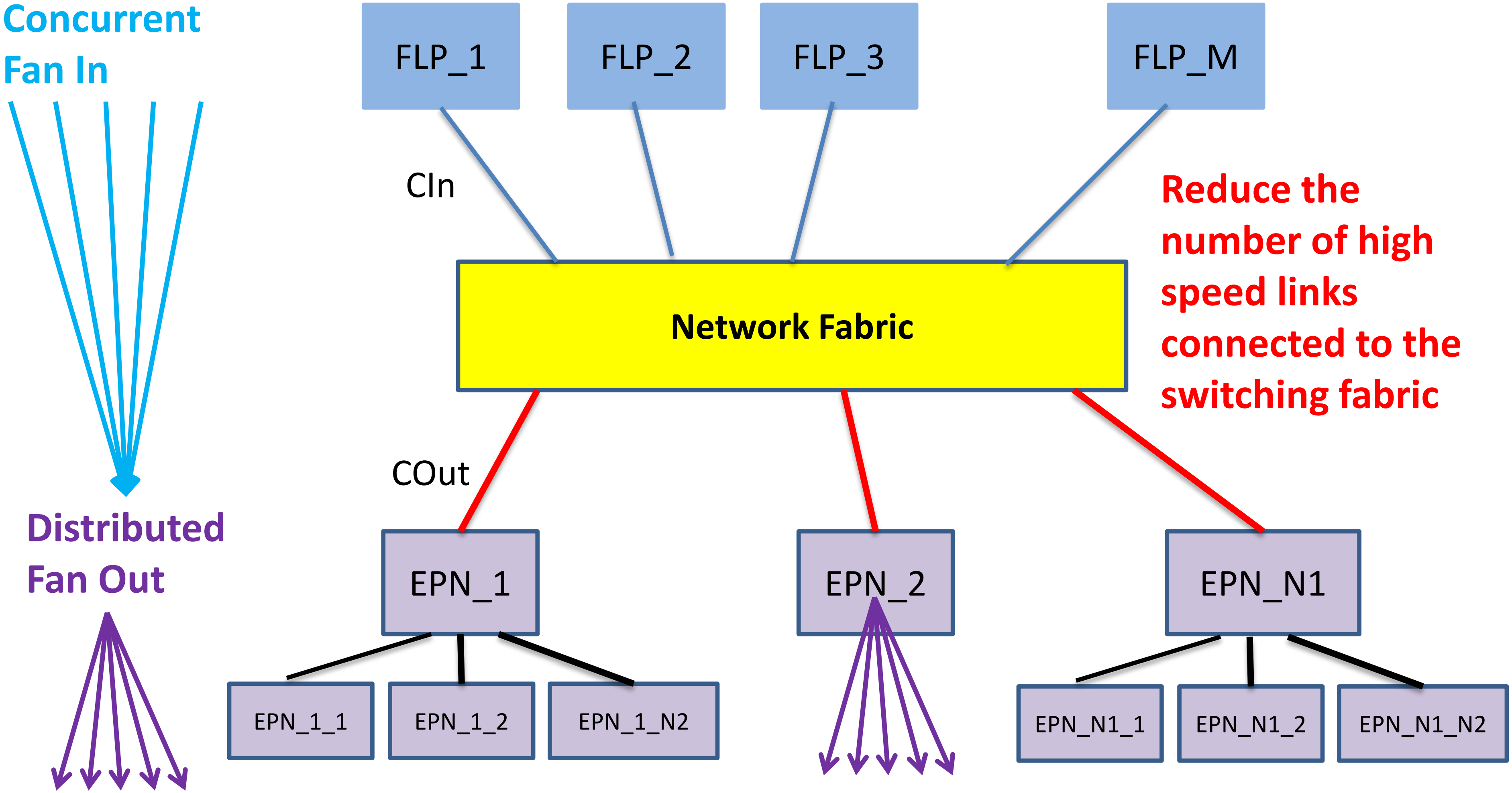


256 FLPs

Two Layer spline leaf
Switch fabric

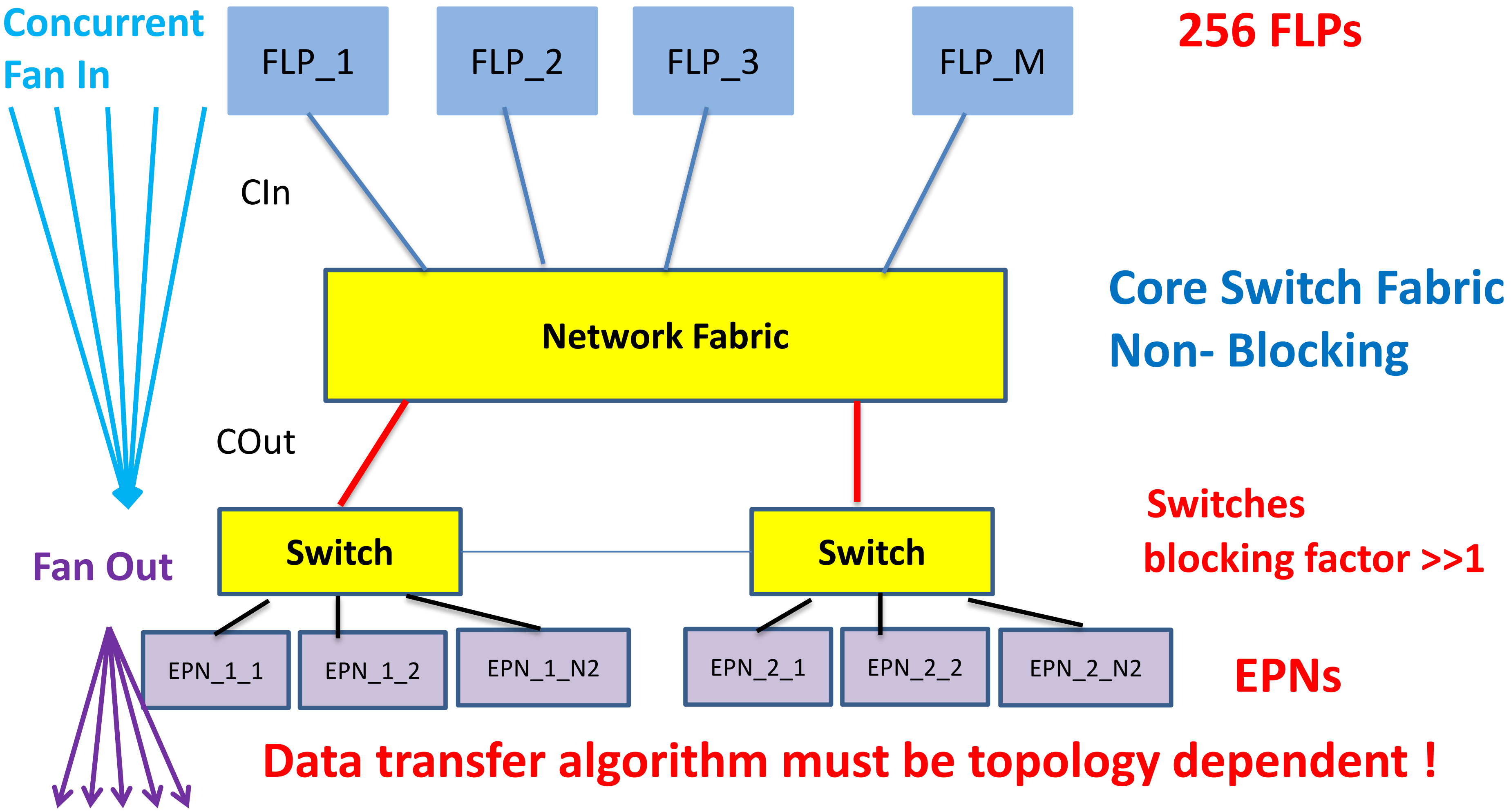
~ 1500 EPNs

FLP – EPN Topology 2



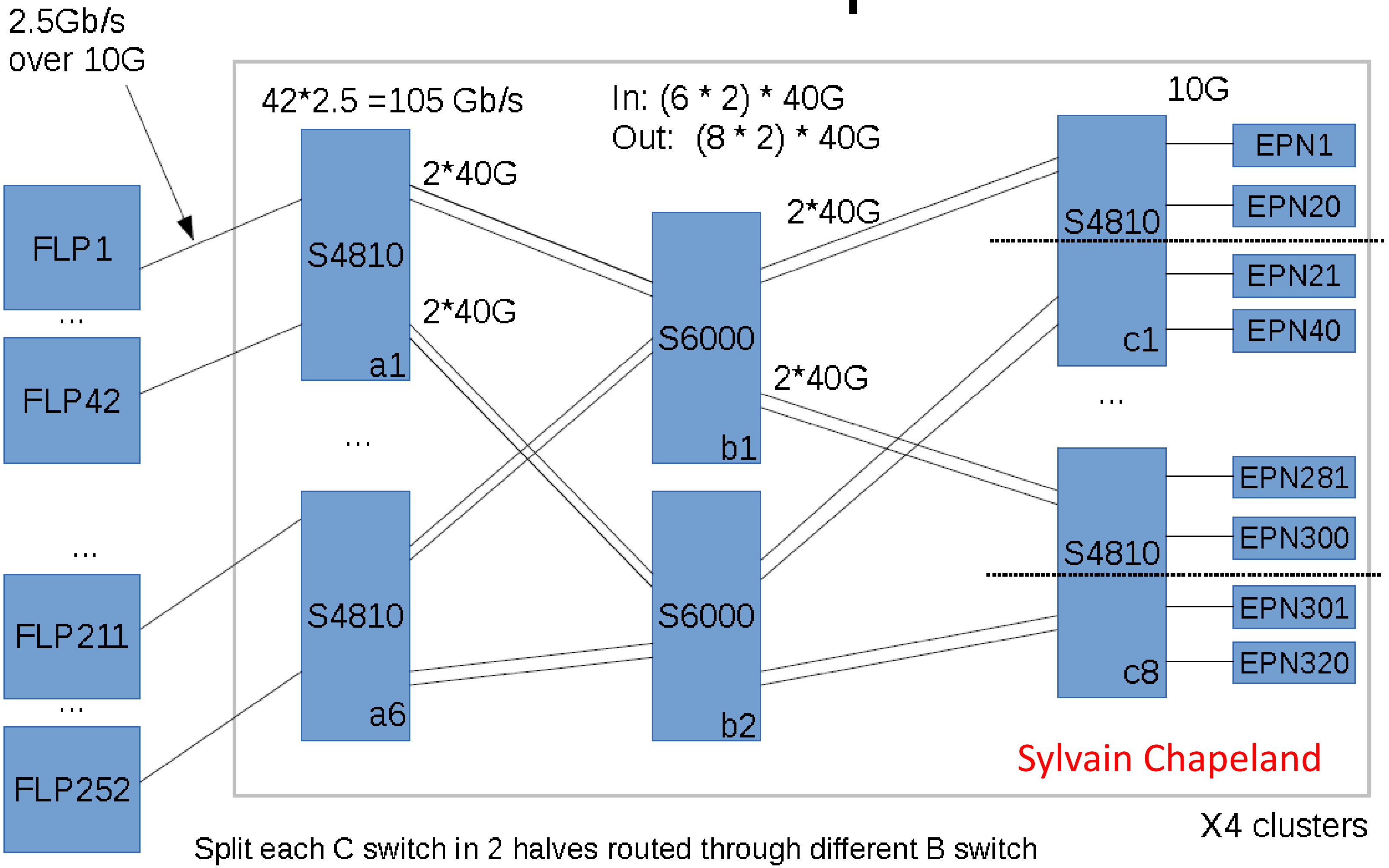
Reduce the number of high speed links connected to the switching fabric

FLP – EPN Topology 3



Data transfer algorithm must be topology dependent !

Cluster fan-out – expendable

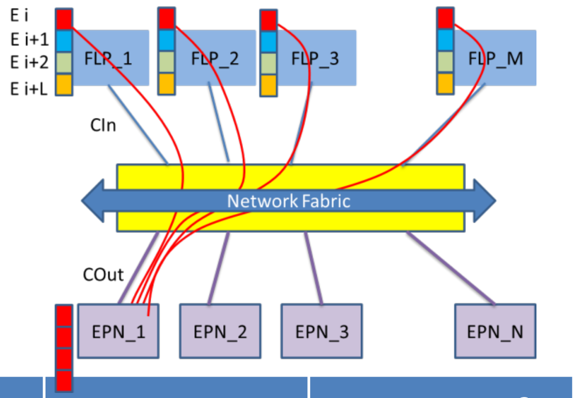


X4 clusters

Split each C switch in 2 halves routed through different B switch
Distribute load so that 8 EPNs active per end-switch (4 EPN per halves)
i.e. 64EPNs active ok: 2 x 4 x 10G < 2 * 2 * 40G
Spare ports: 6*6*10G on A, 2*4*40G on B, 8*8*10G on C. Spare out bandwidth: ~50%

Estimation for the number of concurrent IO processes

data_framgment ~100 MB
250 FLPs
Each EPN receives ~ 10GB
In parallel from all FLPs
50 ms rate for data frames
Bisection Bandwidth ~ 2.5 Tbps



Cout	Cin	Min Latency	Buffer/FLP	Min No Parallel Transfers/FLP	Min No of Concurrent IO processes
10Gbps	10 Gbps	8s	32 GB	250	42 000
20 Gbps	10 Gbps	4s	16 GB	250 (*)	42 000
40 Gbps	10 Gbps	2s	8 GB	250 (*)	42 000
56 Gbps	56 Gbps	~1.5s	5.8GB	60	15 000

We should simulate tens of thousands of concurrent "processes" sending and receiving data

Simulation of IO Processes

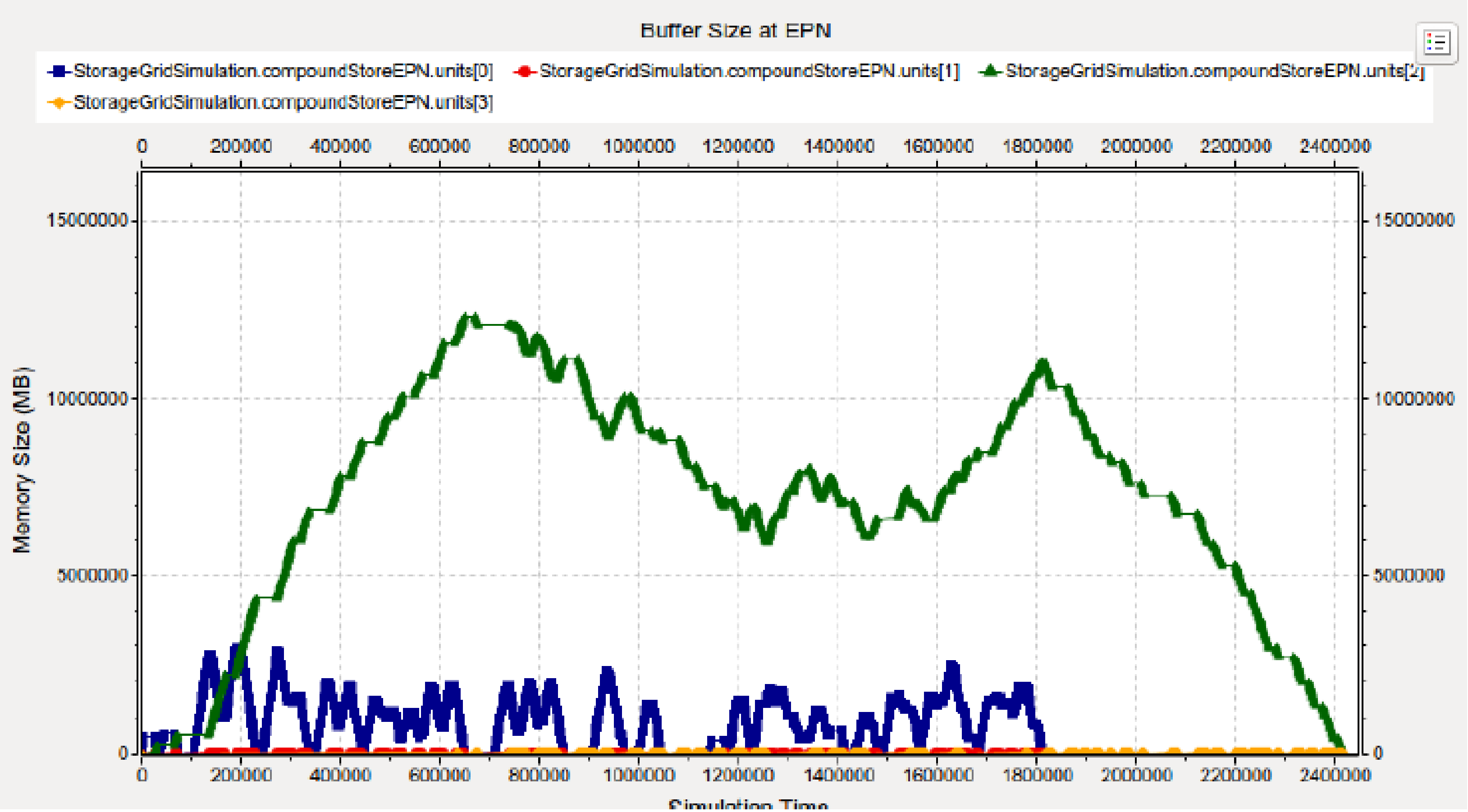
We need to simulate a large number of processes that transfer large amounts of data with constrains.

- **To evaluate different algorithms for data flows, control, error recovery ….**
- **Evaluate the scalability of the system**

Options for simulating interacting programs :

- **Discrete Event OMNet++**
  packet / frame level simulation … may take long time to simulate
- **Discrete Event Process Oriented Simulation (threads - actors) - MONARC simulation tool**
  continuous flow as long as nothing is changing in the system.

OMNeT++ simulation example of Storage



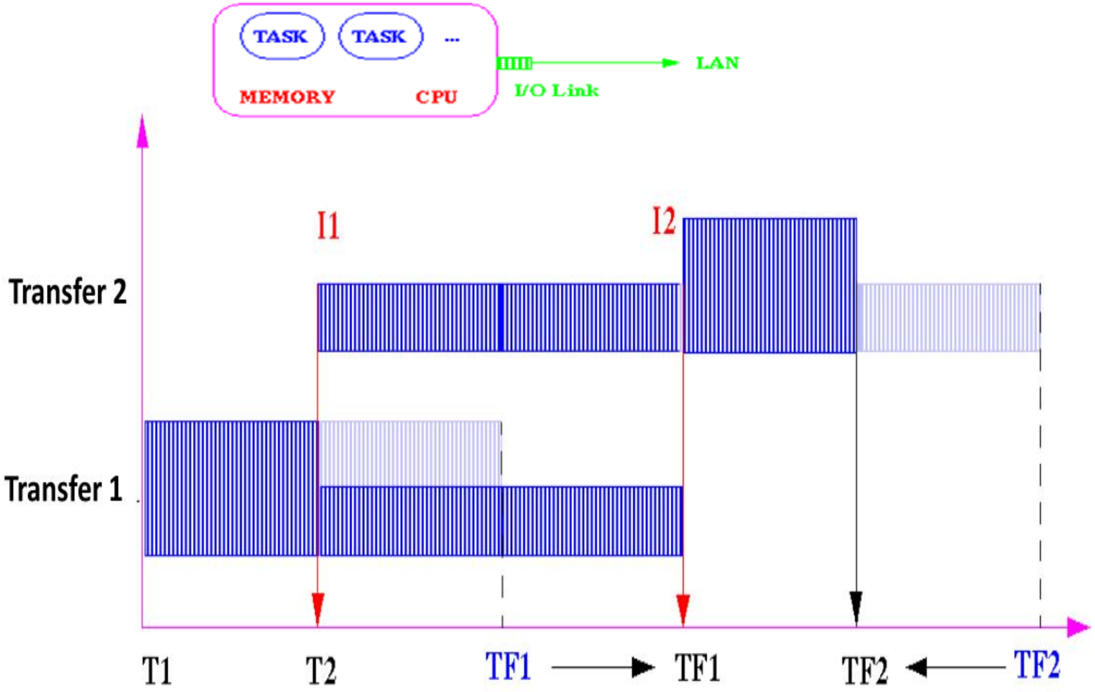
Work started by Charles Delort and is now developed by Rifki Sadikin

Data Transfer and Multitasking Processing Models

Concurrent running tasks (or data transfer jobs) share resources (CPU, memory, I/O links)

"Interrupt" driven scheme: he

For each new task or when one task is finished, an interrupt is generated and all "processing times" are recomputed.



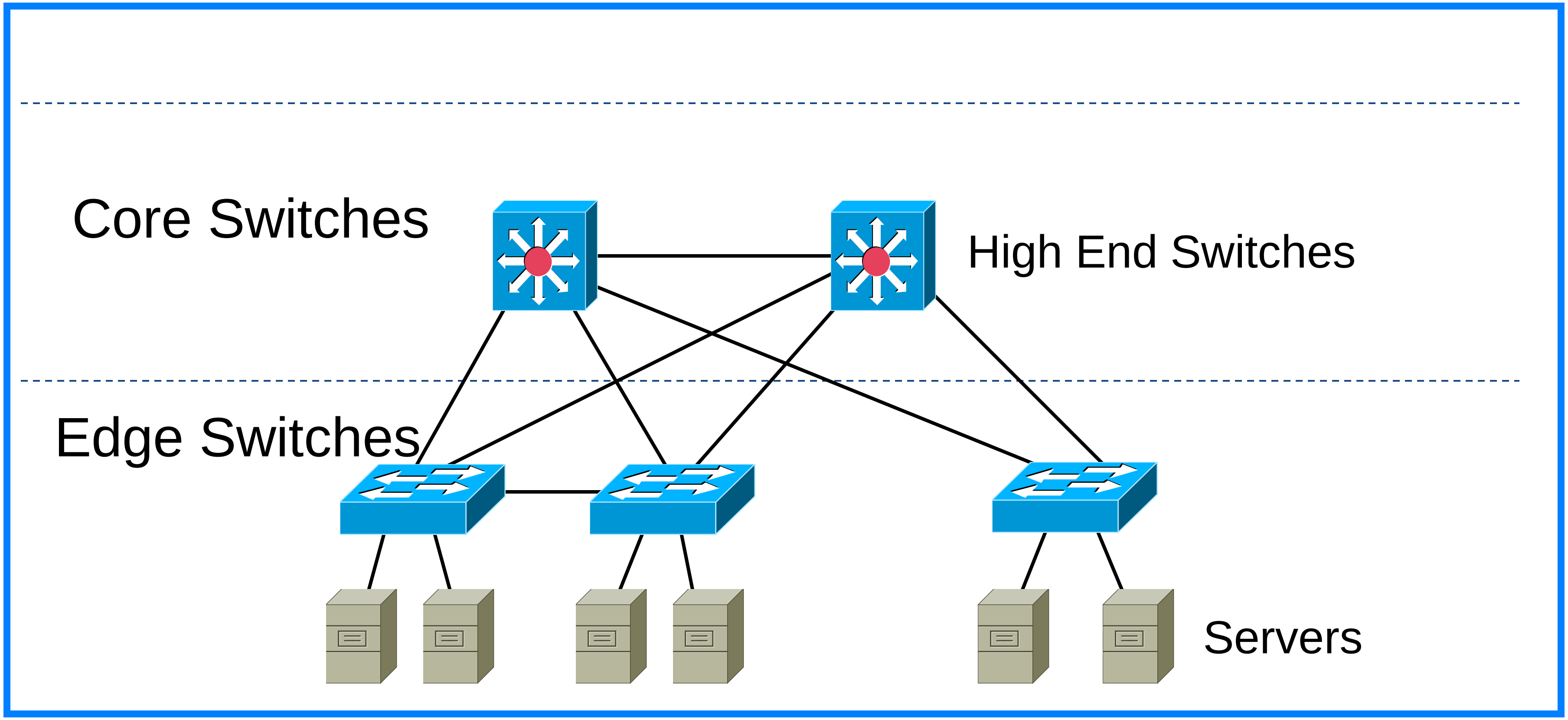
It provides:

An efficient mechanism to simulate multitask processing and **continuous flows**

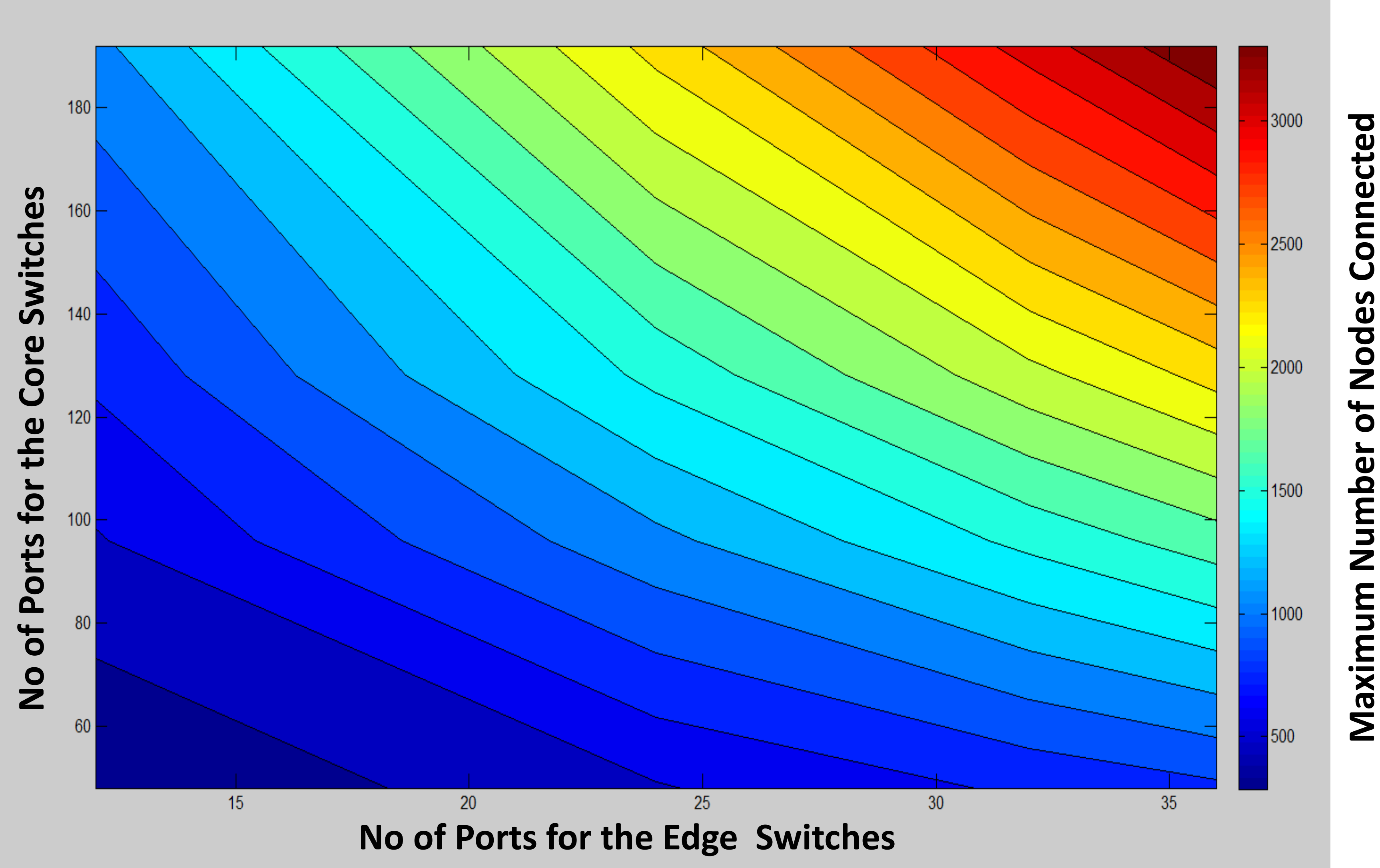
Handling of concurrent jobs with different priorities.

An easy way to apply different load balancing schemes.

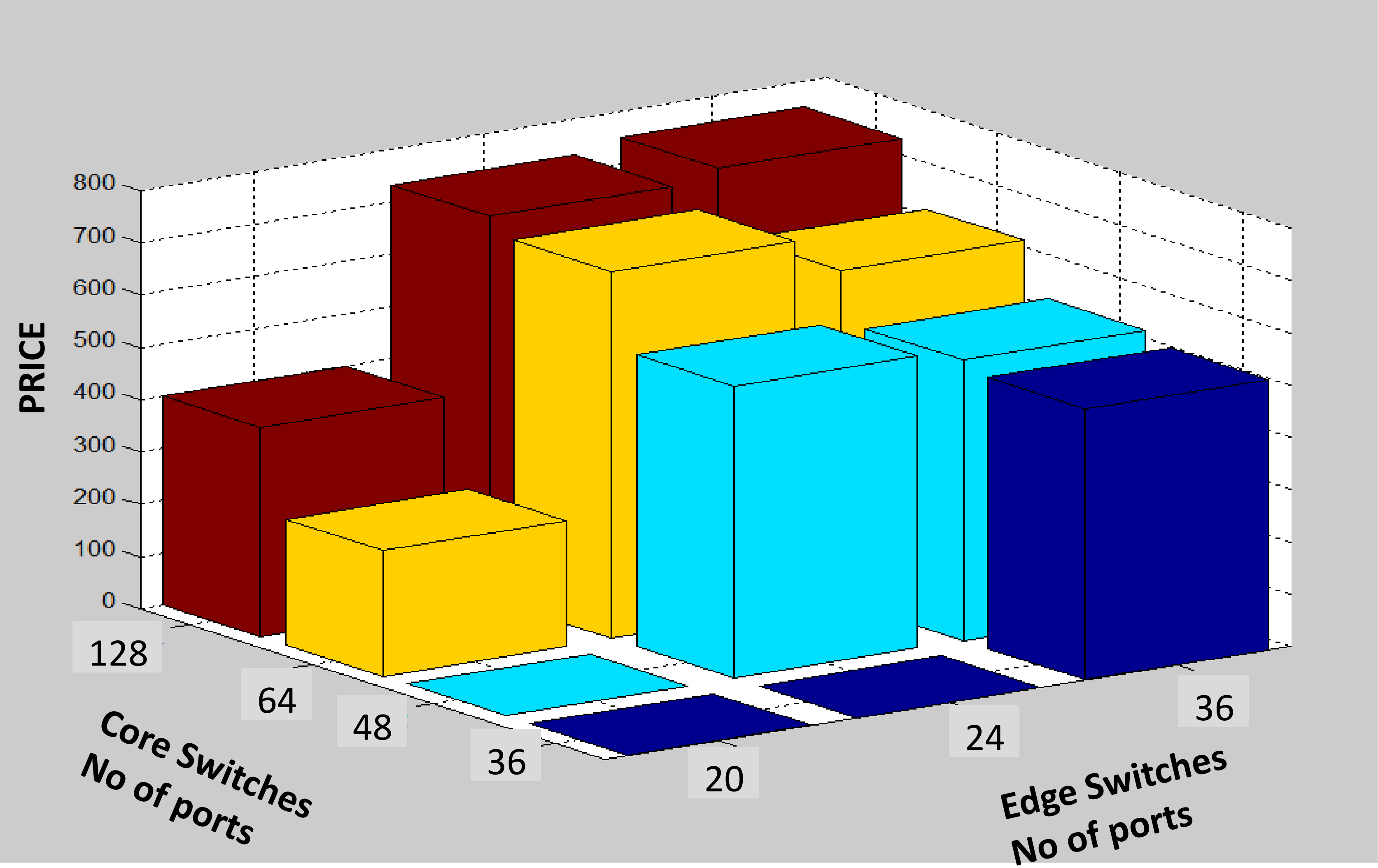
Two Layer Topology
Switch Design



Maximum number of connected nodes for a two layers system (non-blocking)
Select the right technology



Price Example o connect 500 nodes – non-blocking with different switching systems



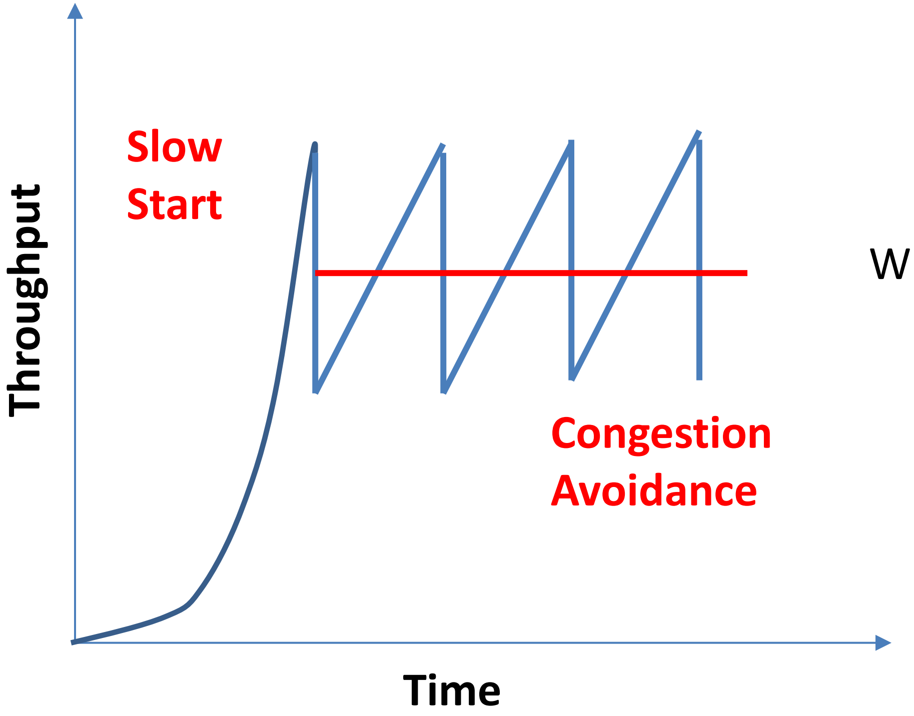
Networking : Transport Layer

Provide *logical communication* between application processes running on different hosts ***The transport layer is responsible for process-to-process delivery***

- **Datagram messaging service (UDP)** ***It does not add anything to the services of IP except to provide process-to-process communication instead of host-to-host communication.***

- **Reliable, in-order delivery (TCP)**
  - **Connection set-up**
  - **Discarding of corrupted packets**
  - **Retransmission of lost packets**
  - **Flow control**
  - **Congestion control**
- **RDMA** the network adapter is capable to transfer data directly to or from application memory

TCP Performance

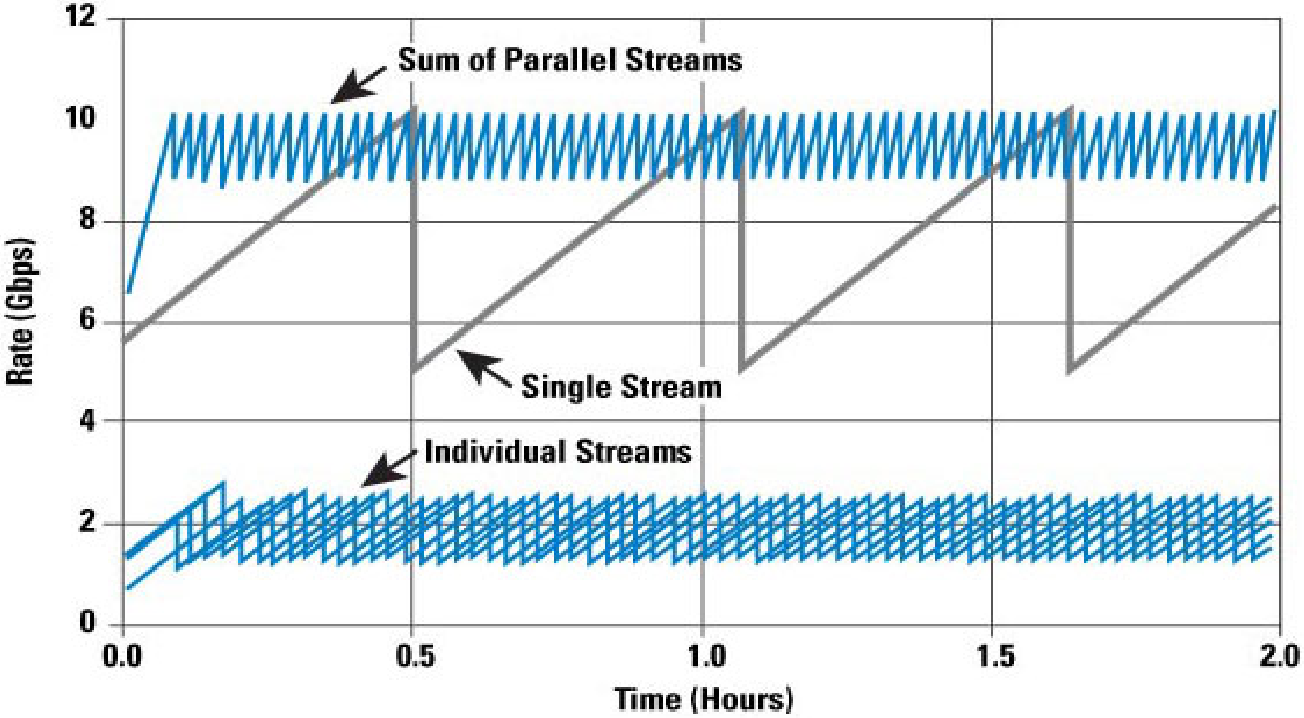


$$BW \sim \frac{\text{Segment Size}}{\text{RTT} * \text{SQRT (Loss Prob)}}$$

What influences the TCP performance?

- **Available bandwidth**
- **Packet Loss**
- **Out of order delivery**
- **Round-trip**
- **Congestion avoidance algorithm**
- **TCP setup and tuning**
- **Buffers in Switches and routers**

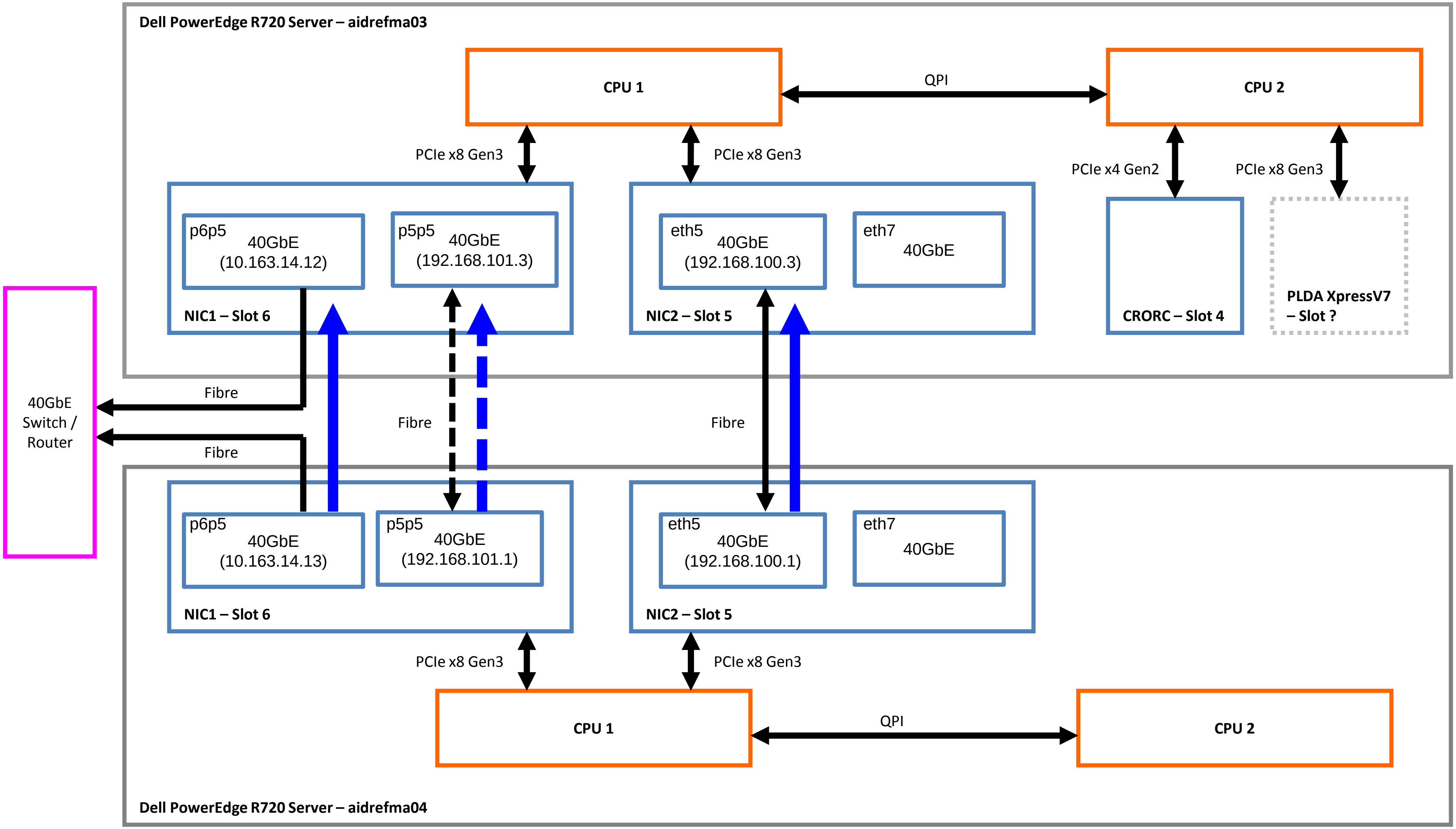
TCP-Tuning for high performance data transfers



- ❑ Significantly increase memory buffers
- ❑ MTU –Maximum Transfer Unit Jumbo frames MTU 9000
- ❑ IRQ pinning (also know as IRQ affinity)
- ❑ Congestion control (cubic)

Network tuning at 10 Gbps is not the same as for 40 Gbps

FLP Memory Bandwidth Tests – Dell PowerEdge R720 Server



Ervin Dénes, Ernő Dávid

FLP 40GbE Bandwidth Tests – Very Preliminary Results

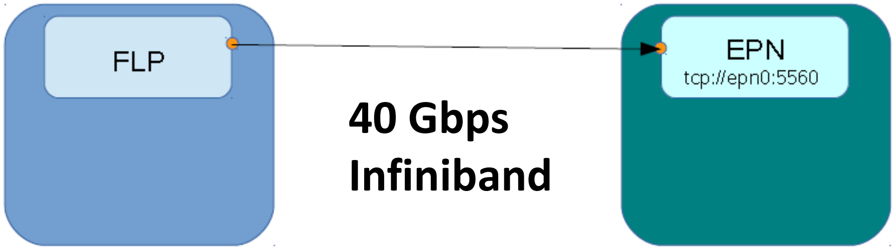
Link	FDT [Gb/s]	iperf3 [Gb/s]	Custom [Gb/s]
p6p1	39.6	19.0	27.5
p6p2	37.5	20.4	11.5
p4p2	33.8	19.9	15.6

OS: CentOS 7, Kernel: 3.10.0, Test: TCP/IP, 4 thread per link

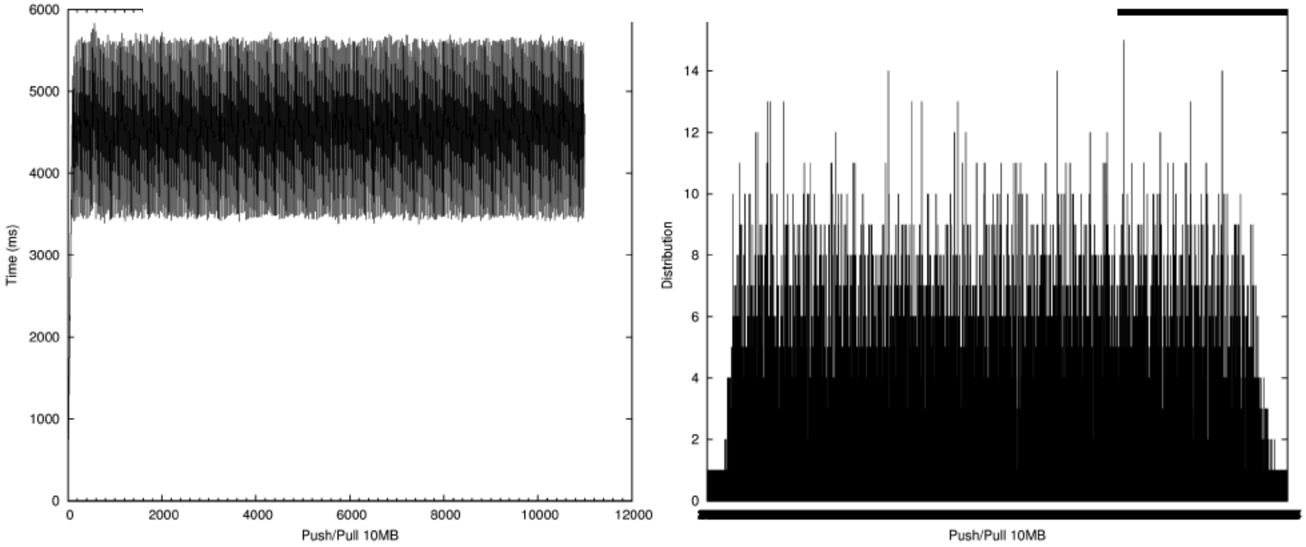
Running all three links in parallel ~85 Gbps aggregate throughput

We can get 100 Gbps throughput with appropriate tuning for the IRQ affinity

ZMQ (ALFA framework) – Performance Tests



40 Gbps Infiniband



Push/pull pattern 10 MB Payload
Transfer rate: ~0-3500 MB/s

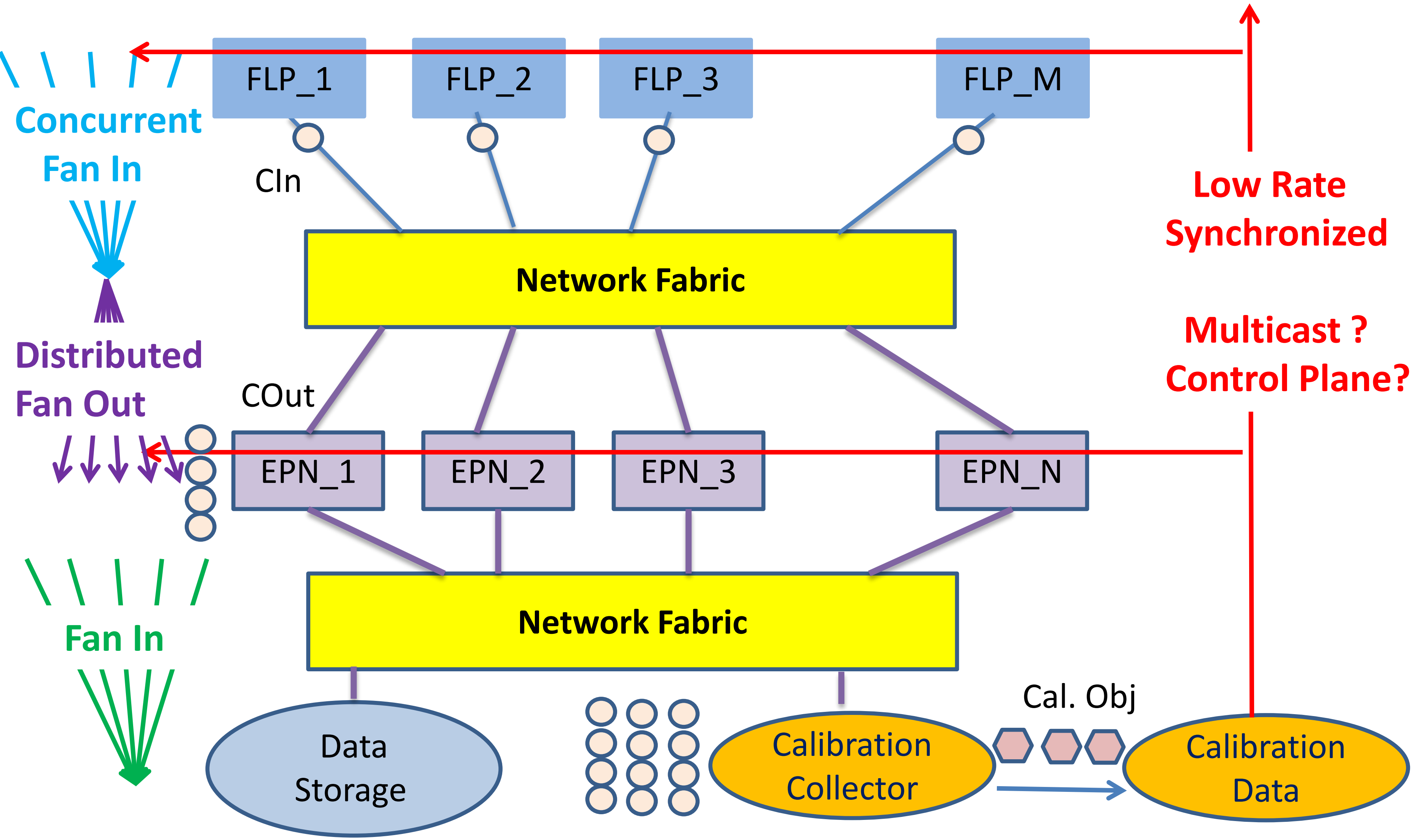
Charalampos Kouzinopoulos
Mohammad Al-Turany

Topology Optimization

- **Based on existing technologies evaluate possible topologies that can perform the task.**
  - **How much it can scale …**
  - **Risk analyses (performance degradation if one or several switches fail)**
  - **Inefficiencies in resource utilization**

- **Based on the price estimates, evaluate a subset of effective topologies**
  - **Select the adapted algorithms for the data transfer control for each of these topologies**
  - **Technology evolution vs price ?**

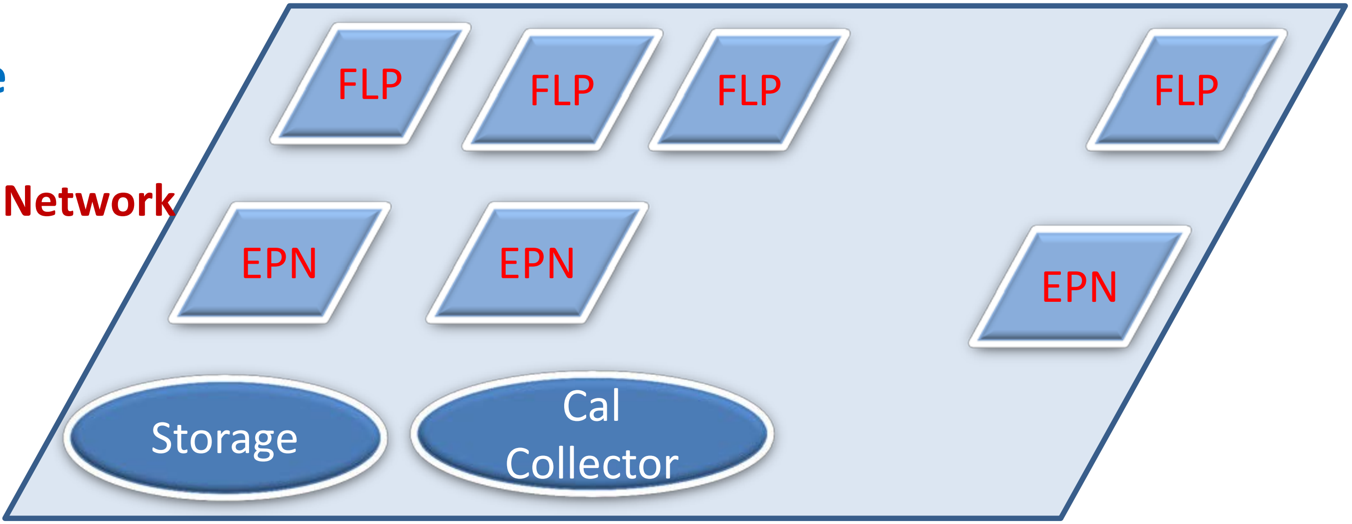
The Calibration Data Traffic



Data and Control Plane

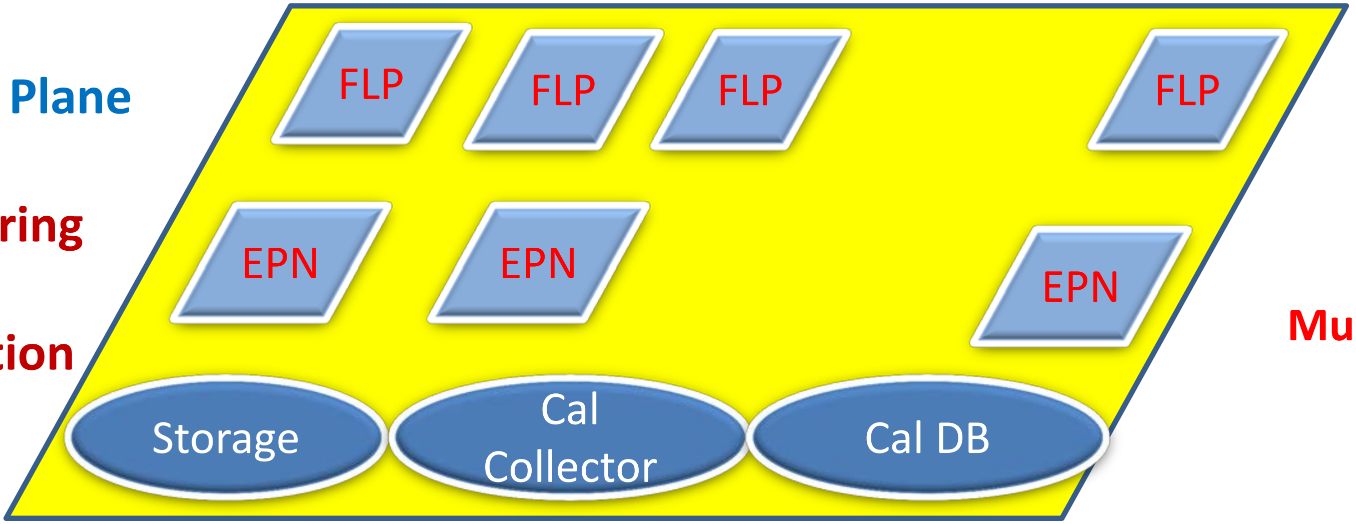
Data Plane

High Speed Network



Control Plane

Monitoring
Control
Calibration
....



Multicast ?

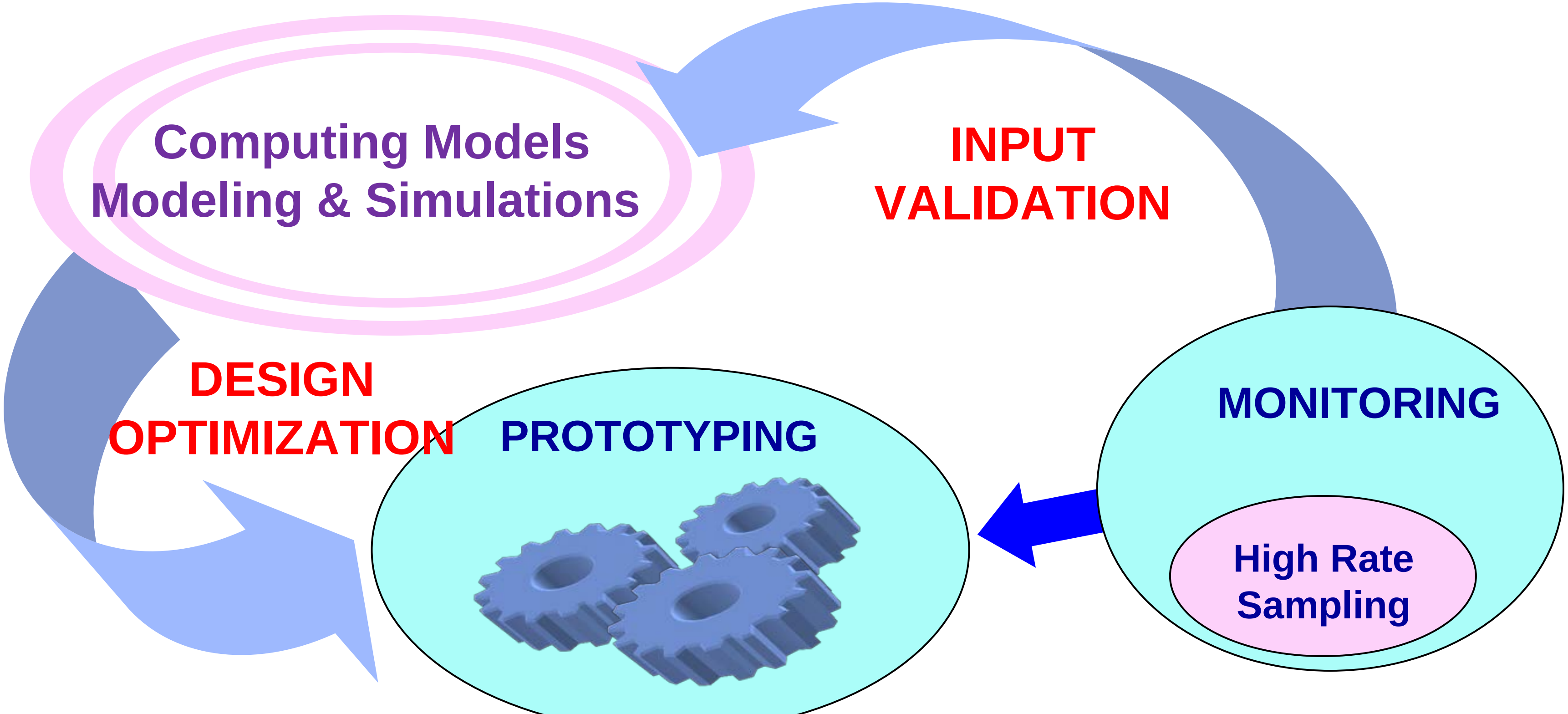
Data Flow for calibration

- FLP - > EPN traffic   addition of small calibration data structures

- EPNs collect these structures and sends them to a Calibration data collector . Than it generates the calibration data objects

- The calibration data objects should be synchronously distributed to all FLPs and EPNs units (at a low rate)

Simulation, Modeling and Monitoring are essential for an efficient, cost-effective computing system



Simulation and modelling should be part of the system design as continuous process to validate and optimize the overall computing model. Computing system simulation should be something very similar with Monte Carlo simulations for the physics part.

Summary

- ✓ Continue the simulation work for the main three possible architectures
- ✓ Collect realistic data for price estimation of different technologies and switches
- ✓ Perform test bed measurements and estimate the performance of different data transfer software. IO tuning Include these values into the simulation
- ✓ Define realistic estimate for the calibration flow (together with CWG13) and include these flows into simulation

Close collaboration with the other O2 working groups

Thank you !

Questions ?