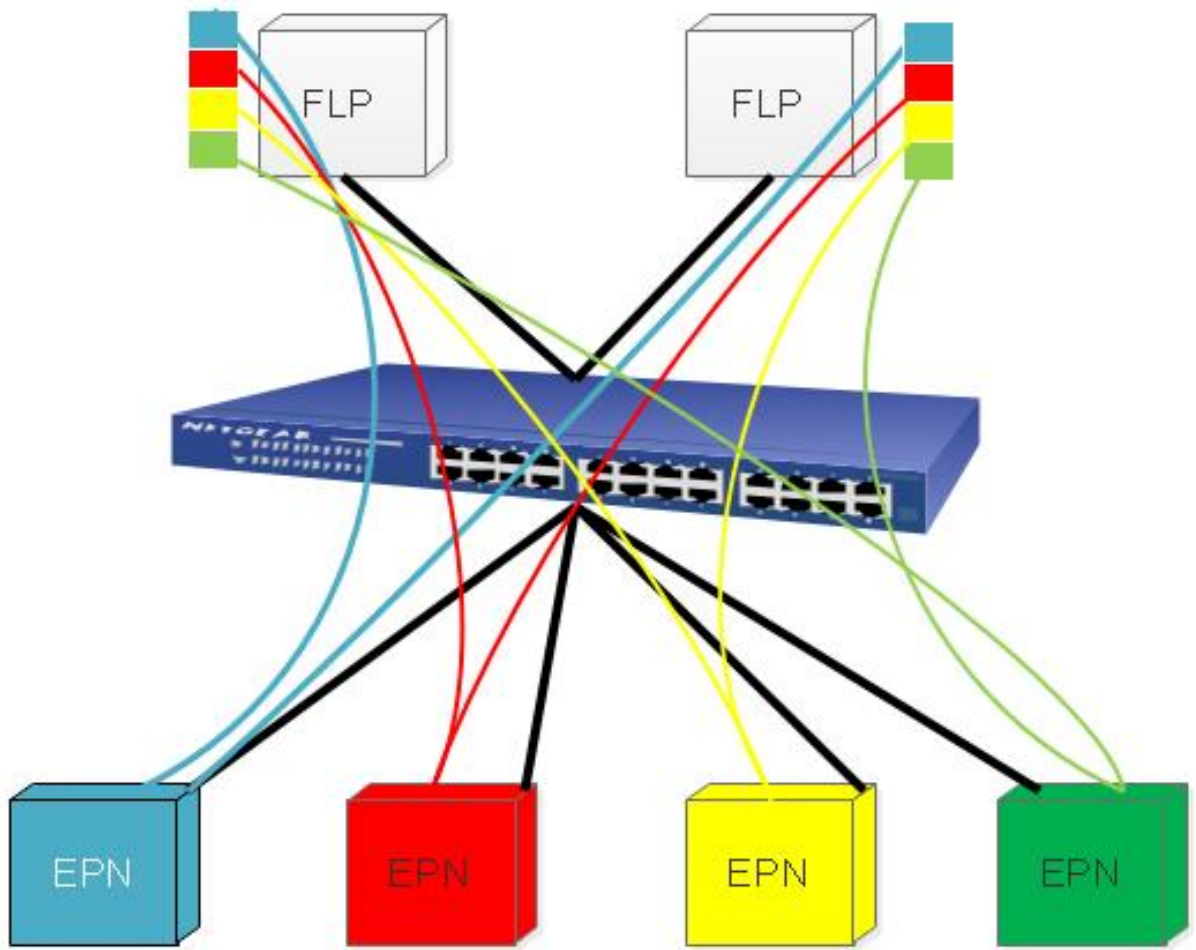
ALICE O2 WP2 Group

Professor Syed Asad Hussain
Dean Faculty of Information Sciences and Technology
COMSATS University Islamabad, Pakistan

WP2-Data Flow and System Simulation

- Headed by Dr. Iosif Legrand at CERN
- WP2 consists of three deliverables
  a) Develop a tool for detailed simulation of network topologies for O2 facility for a quick evaluation of multiple scenarios
  b) Develop a tool to allow for a quick estimate of computing resource needs
  c) Estimate overall long term computing needs for future runs and Grid resources

Simulation Scenario



WP2-Data Flow and System Simulation Introduction

- Constitute a system that handles data transfer at high speed of atleast 4Tbps
- Basic scenario constitute of 250 FLPs (low processing nodes) and 1500 EPNs (high processing nodes)
- Data Transmission consists of creating full time frame building tasks where the task of building the full time frame is to send the time-frame data chunks from all FLPs to one EPN unit.
- Each FLP is supposed to generate 40MB @ 50Hz (every 0.02 seconds)
- Total input rate at a FLP is ~500Gb/s and the average bisection traffic through the switching fabric is ~4Tb/s
- EPN Input speed required 400 logical connections to send all 4Tb to one EPN

Introduction (Cont..)

- The link rate between FLP and switch is 40Gbps and from switch to EPN is 10Gbps
- TCP is being used as the primary transport protocol
- Simulation at COMSATS Lahore lab is done using NS-3 due to its flexibility and it being an open source

Some Challenges

The main task which we are handling is (The Full-Time Frame Building), figure # 1 shows the schematic views of the data flow to create complete time frames on the EPN nodes. Following are the main challenges:

- To be able to keep up with the high input rate.
- ~ 250 FLP processing units that collect sub-detector time frames and using a control algorithm send the data to assemble the full-time frame in one EPN node. (we need to avoid congestion on the switch) (TCP throttling).
- Scalability issue (need to design a system that is scalable).

Network Simulation Tools

- NS3 is being used at Lahore Campus and OMNet++ at Islamabad campus.
- A complete customized transfer helper is developed to meet the requirement.
- Maximum MTU, Jumbo Frames and Maximum Packets size is used to test the given scenario at full load.
- Helpers are created in Java, C++ and even Python to see the performance variations on these different platforms.
- High Computing Machines and Servers are dedicated for this task

Simulation Parameters for NS3

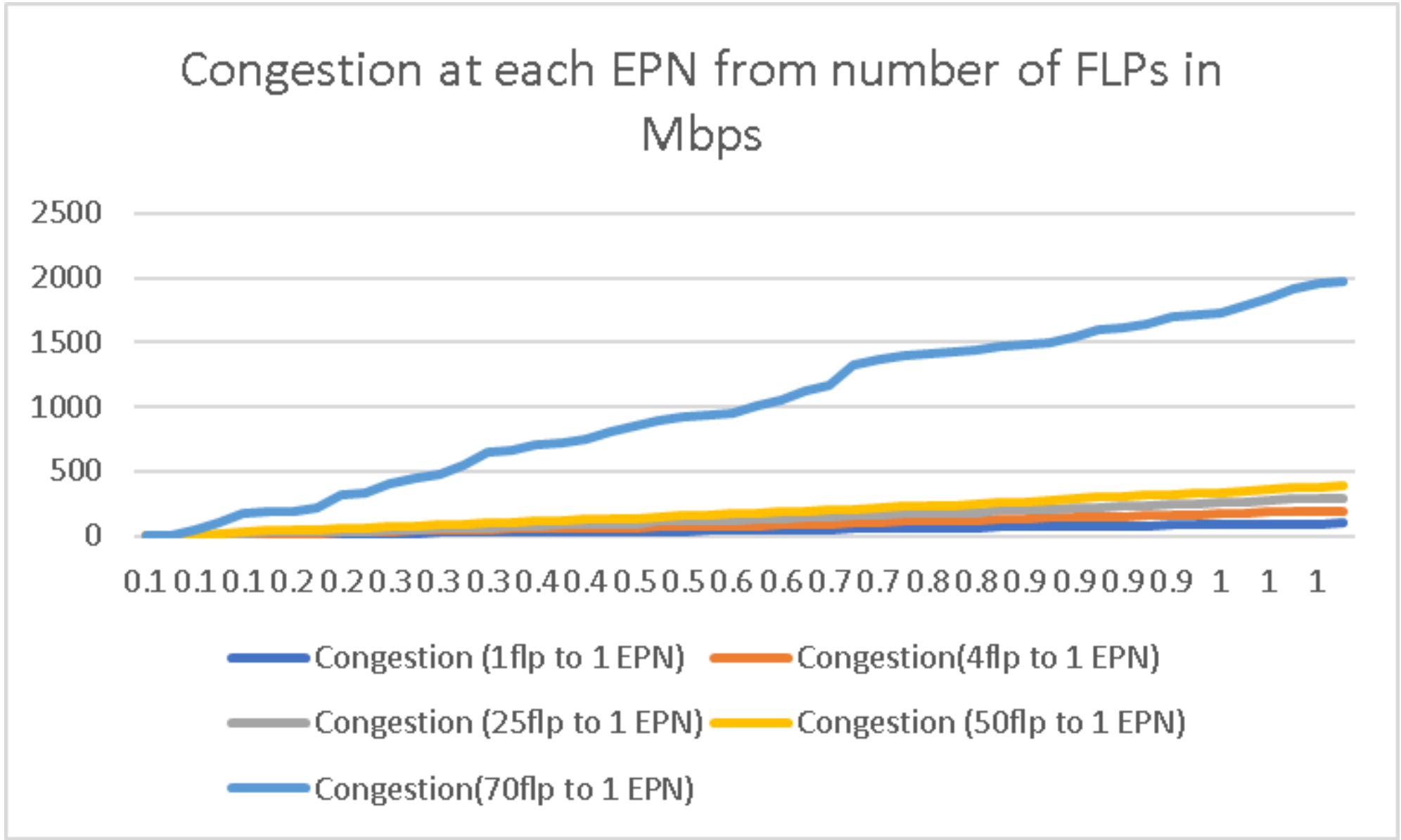
Name	Value (TCP)
No of FLPs	1, 4, 25, 60, 70, 250
No of EPNs	6,24,150,360,420, 1500
Input Rate at FLP	40MB in 50Hz
FLP-Switch Link Rate	40Gbps
Switch-EPN Link Rate	10Gbps
Simulation Time	4 hours
Simulator	NS-3

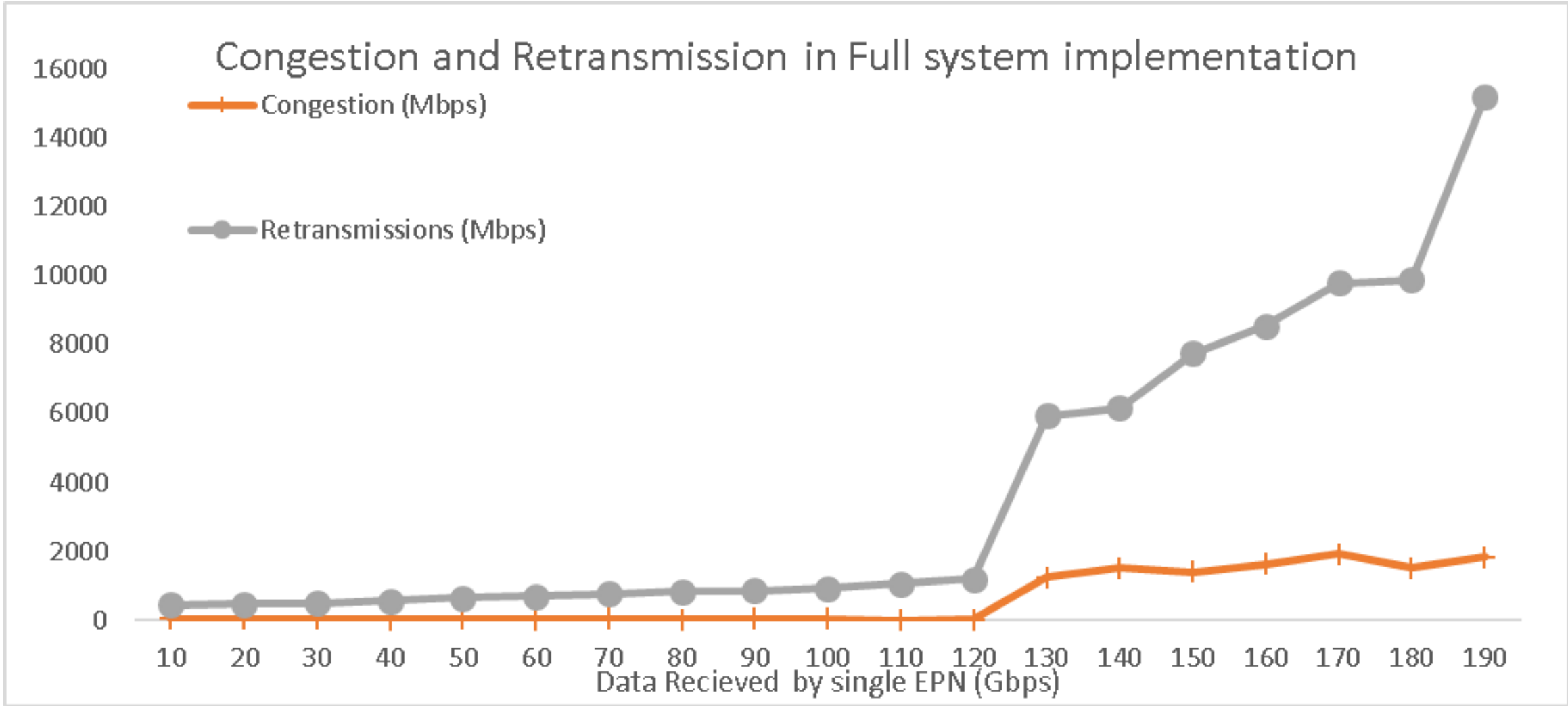
Problems faced so far

- Each node/EPN based on its characteristics and input rate accepts only 130Gb of data without congestion and retransmission attempts.
- Using different TCP variants, the congestion and retransmission has reduced the simulation time from 24 hours to 5 mins for 5 FLPs to 30 EPNs.
- A full scale system (250 FLPs and 1500 EPNs) in default configuration of NS3 took more than 5 days and we reduced this time to 24 hours by developing custom helper and than 4 hours.
- Default TCP congestion mechanism are not suitable for simulating this full scale system.

Results (Default TCP configuration Mechanism) (initially)

No. of FLPs	Throughput at each EPN(Gbps)	Time taken for full system run	Congestion (Mbps)
1	10	2s	97
4	10	1 min	100
25	9.99	5 mins	98
60	9.99	8 mins	97
70	9.28	45 mins	1600


Congestion at each EPN from number of FLPs in Mbps
2500
2000
1500
1000
500
0
0.1 0.1 0.1 0.2 0.2 0.3 0.3 0.3 0.4 0.4 0.5 0.5 0.6 0.6 0.7 0.7 0.8 0.8 0.9 0.9 0.9 0.9 1 1 1
Congestion (1flp to 1 EPN)
Congestion(4flp to 1 EPN)
Congestion (25flp to 1 EPN)
Congestion (50flp to 1 EPN)
Congestion(70flp to 1 EPN)

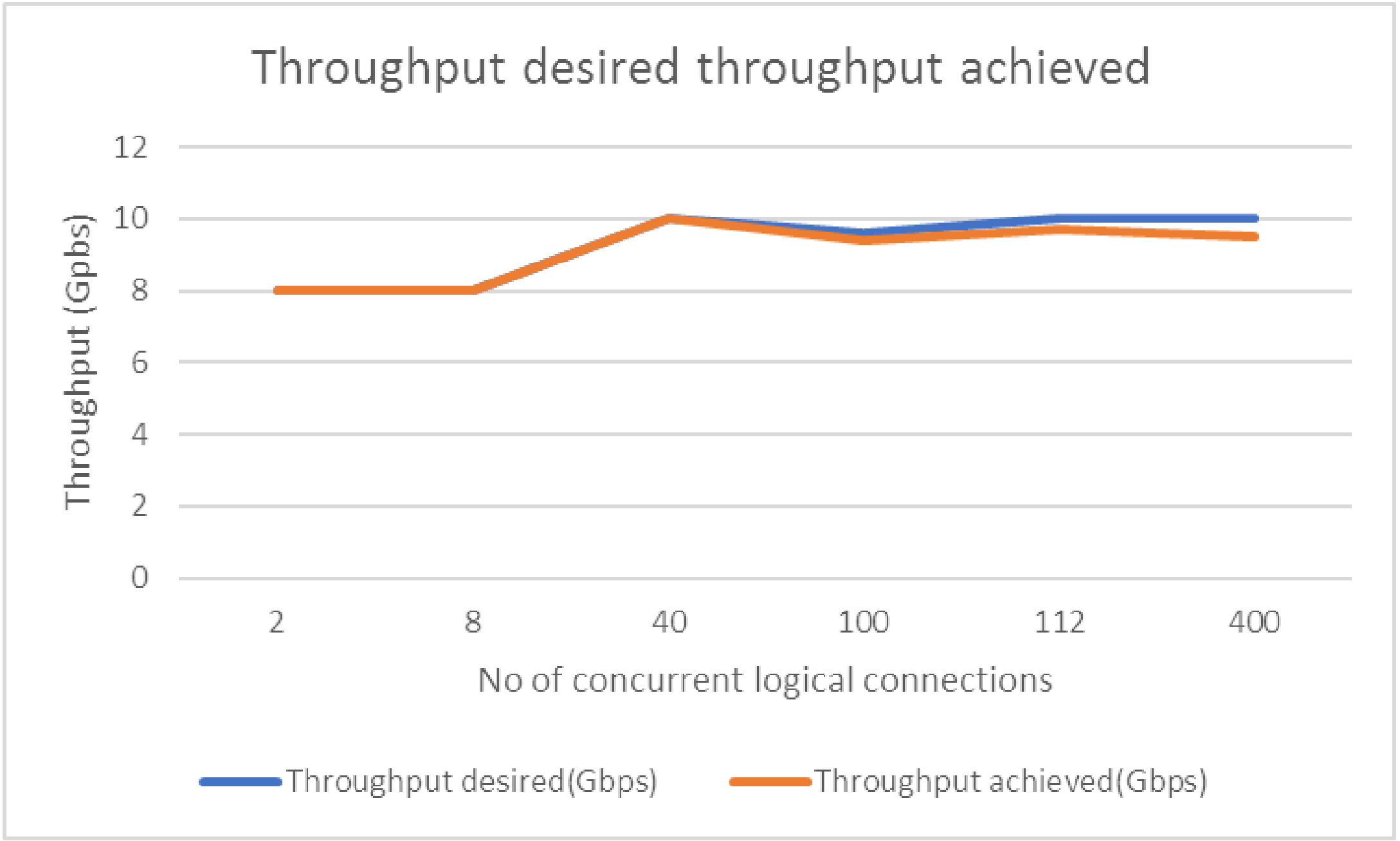

Congestion and Retransmission in Full system implementation
Congestion (Mbps)
Retransmissions (Mbps)
16000
14000
12000
10000
8000
6000
4000
2000
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
Data Recieved by single EPN (Gbps)

Current Progress

- We are working on different congestion management scenarios in NS3 and looking at custom options too.
- We have successfully tested NS3 to its limits and are now in process of modifying it to full extent to implement the full system i.e. Highest MTU, Largest Packet size and Jumbo frames.
- The full scale system with only 4% retransmissions has been implemented in completion time of **4 hours** generating 4Mb congestion.
- These simulation results are 0.004% different from actual system implemented by Dr. Iosef. Legrand at CERN.

Results

No. of FLPs	No of Logical Connections	Throughput at each EPN(Gbps)	Time taken for full system run	Congestion (kbps)
1	2	8	2s	23
4	8	8	1 min	40
25	40	10	5 mins	120
60	100	9.4	8 mins	415
70	112	9.7	45 mins	1200
250	400	9.5	4 hours	4000


Throughput desired throughput achieved
12
10
8
6
4
2
0
Throughput (Gpbs)
2
8
40
100
112
400
No of concurrent logical connections
Throughput desired(Gbps)
Throughput achieved(Gbps)

Summary

- The full scale system with 250 FLPs and 1500 EPNs have successfully been implemented using NS3
- The full scale simulation took 4 hours to complete
- The congestion has been reduced from 1.6 Gb to 4Mb
- Number of retransmission attempts have been reduced.

Simulation setup in OmNet++

Table 1: Simulation parameters for different network scale size

Parameters	Architecture I	Architecture II	Architecture III
Number of FLPs & EPNs	3 FLPs & 18 EPNs	50 FLPs & 300 EPNs	250 FLPs & 1500 EPNs
Number of switches	Single switch		
Packet size	0.4 GB		
Link capacity	From FLP to Switch 10 Gbps From Switch to EPN 10 Gbps	From FLP to Switch 40 Gbps From Switch to EPN 10 Gbps	
Traffic generation	0.02 seconds		
Simulation time	~2.0 seconds	~1.0 seconds	~3.0 seconds

Simulation Time Vs Real Time

Table 2: Simulation Time Vs real time

Architecture	Real time taken	Simulation Time
Architecture I	4 – 5 hours	~2 sec
Architecture II	~ 25 hours	~1 sec
Architecture III	+ 35 hours	~3 sec

Output file size

Table 3: Output (Result) file size of different Versions

Architecture	Size of Output file (before-Old) (Using GUI)	Size of Output file (after-new) (Using Cmd)
Version I	>300 GB	-
Version II	>350 GB	10 MB approx.
Version III	>500 GB	13 MB approx.

Simulation Results (Throughput)

Table 4: Throughput of different network scale size (SINGLE EPN)

Architecture (FLP-EPN)	**Architecture I (3-18)**	**Architecture II (50-300)**	**Architecture III (250-1500)**
Throughput	0.191 Gbps	3.9 Gbps	9.0 Gbps

Throughput (3 FLPs-18 EPNs)

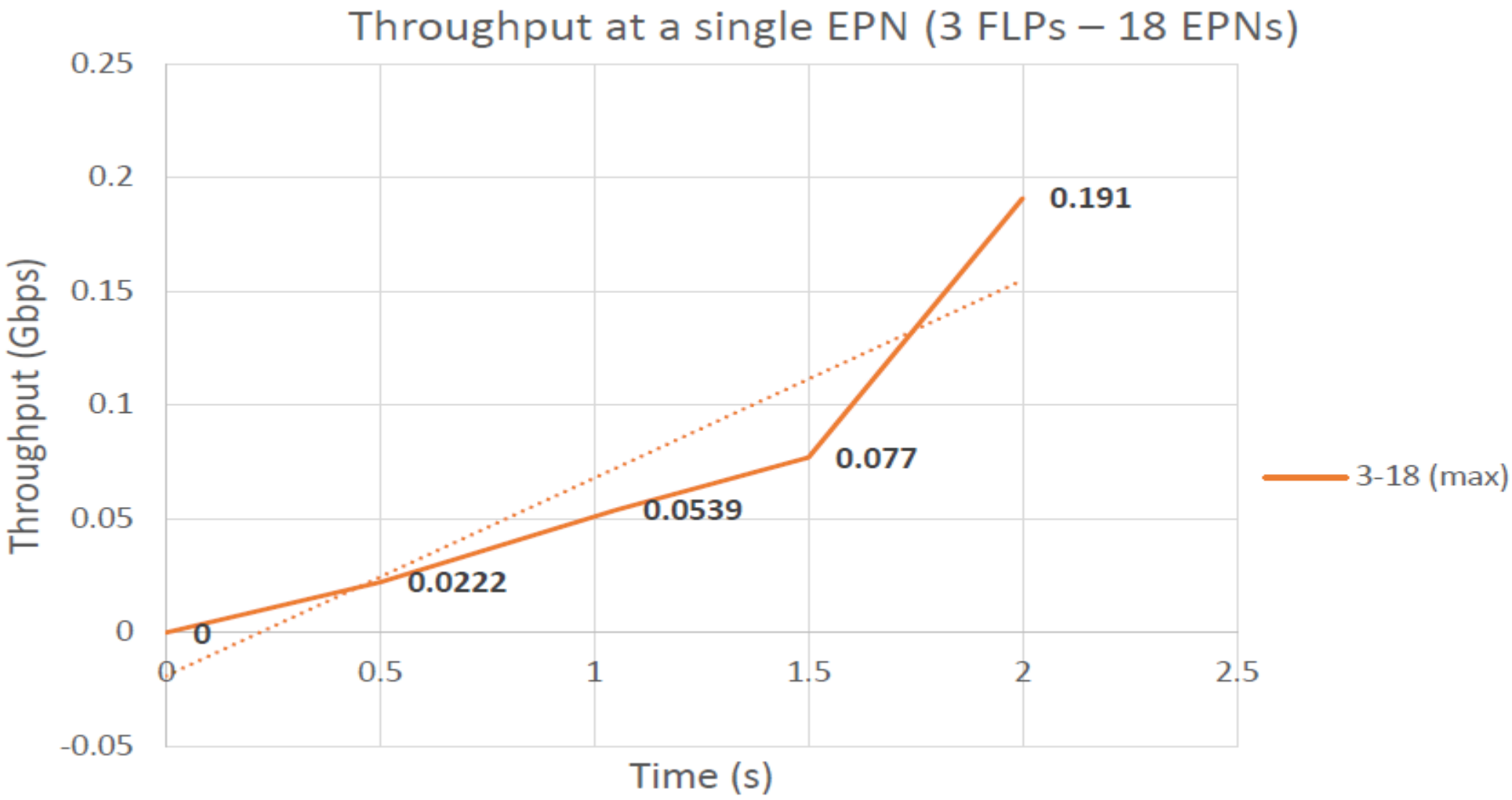


Figure 2: Throughput for network size 3Flps – 18Epns

Throughput (50 FLPs-300 EPNs)

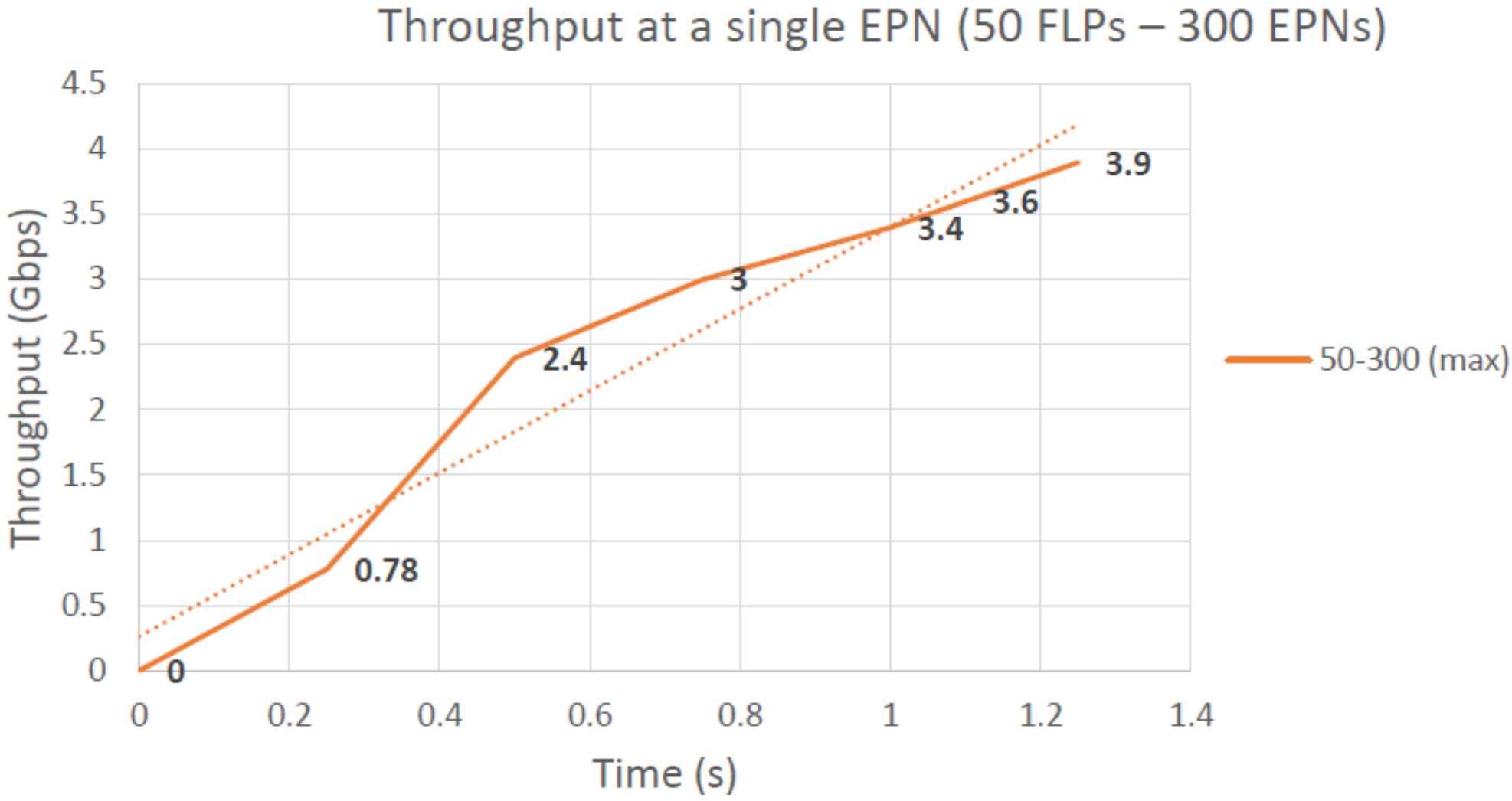


Figure 3: Throughput for network size 50Flps – 300Epns

Throughput (250 FLPs-1500 EPNs)

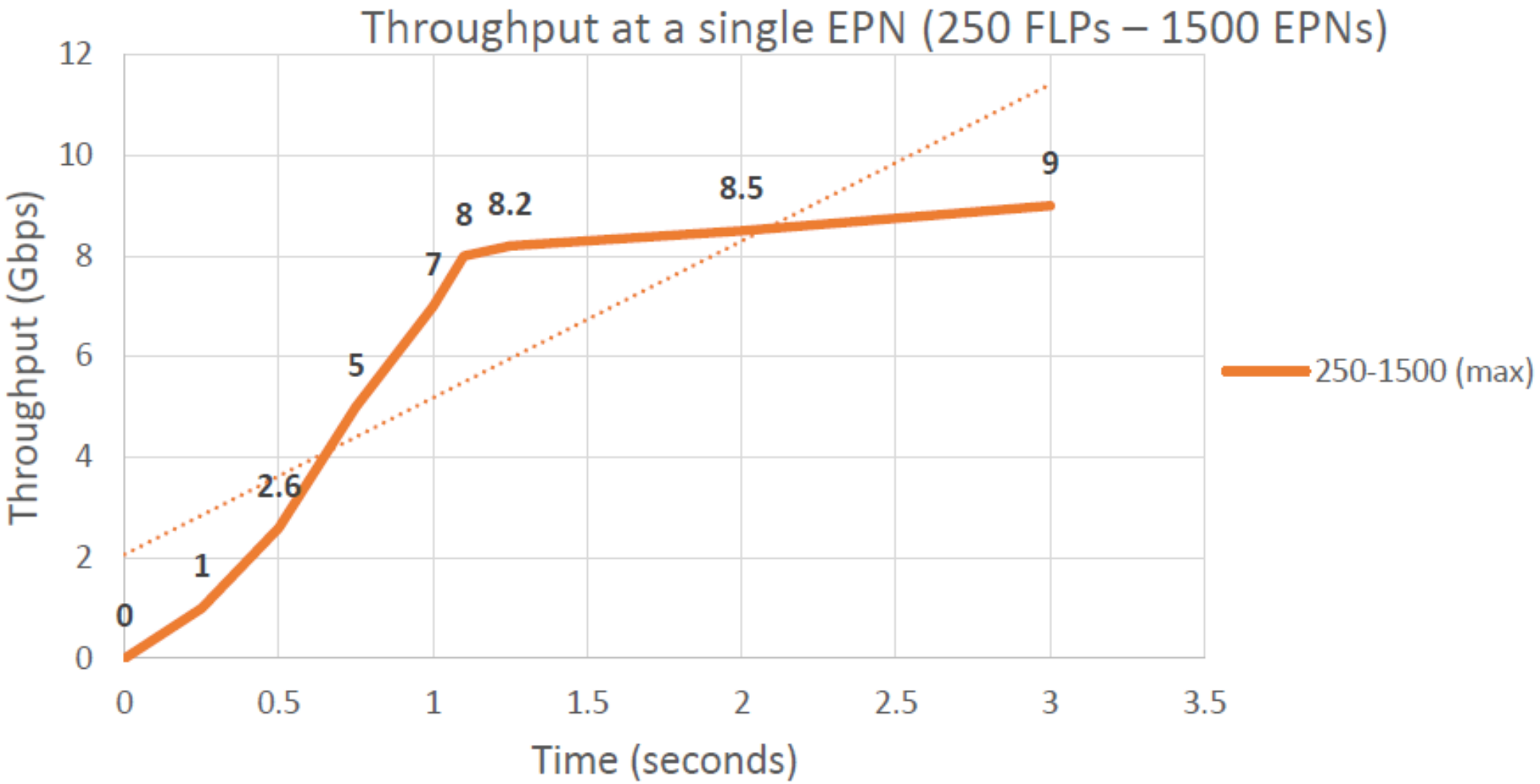


Figure 4: Throughput for network size 250Flps – 1500Epns

More Details:

Architecture(s) proposed by COMSATS Islamabad.

MULTI-LAYERED ARCHITECTURE:

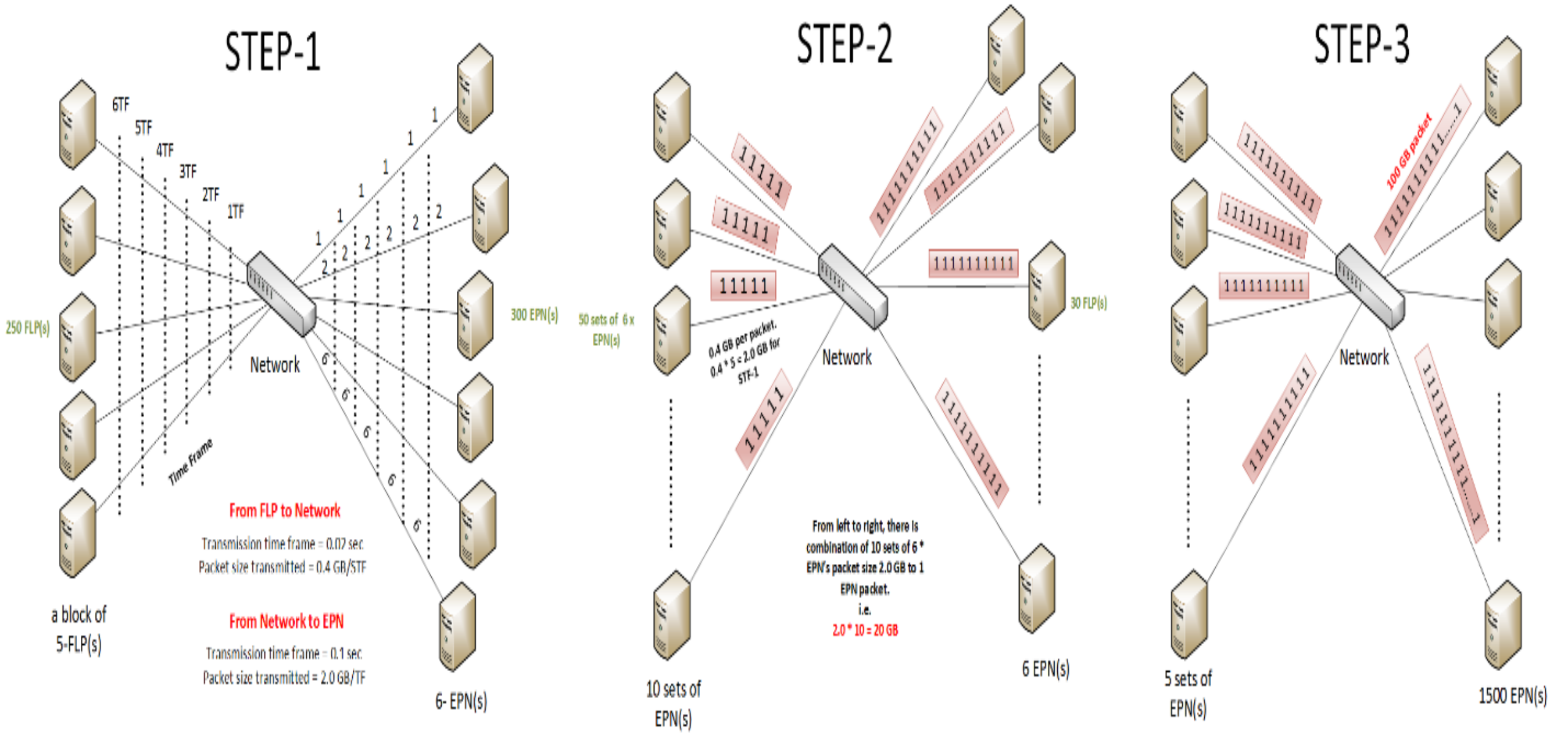


Figure 2: Multi-layered architecture

Congestion control architecture

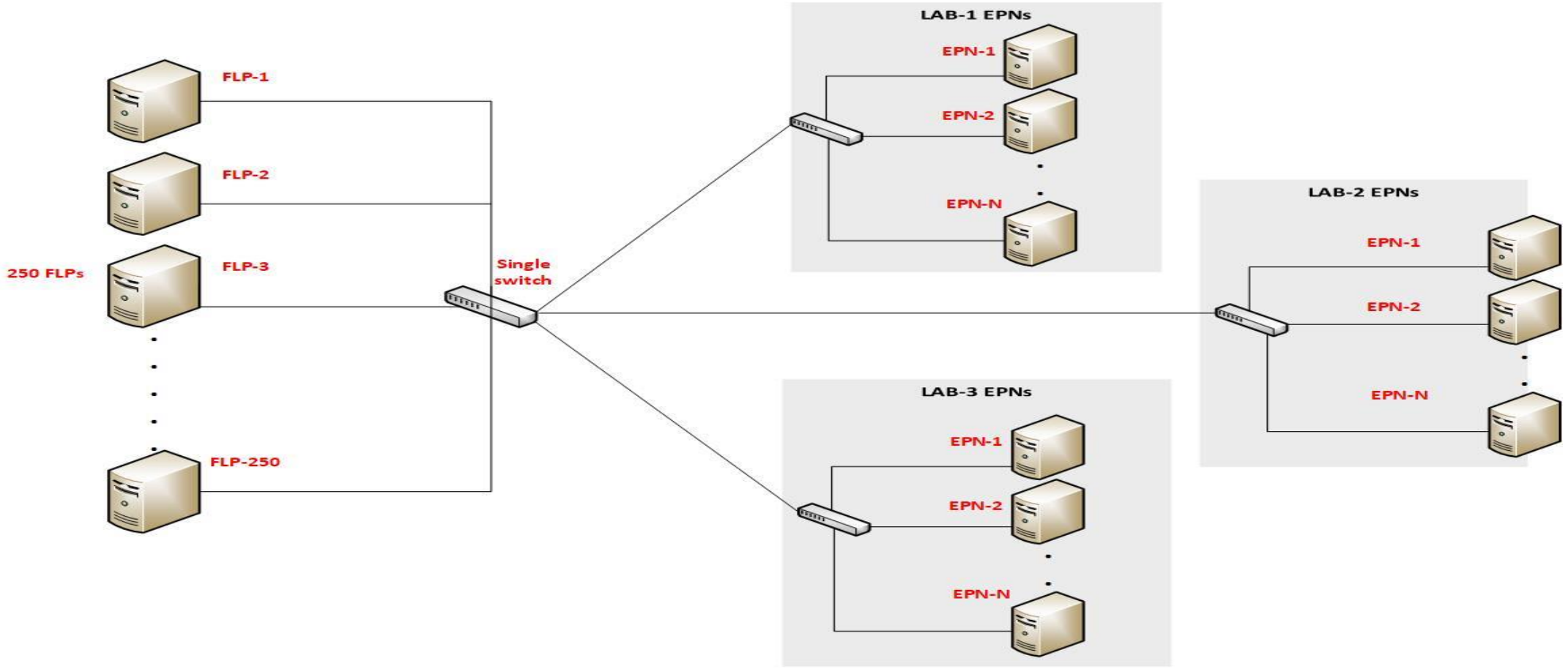


Figure 3: Congestion control mechanism (50 labs i.e. containers)

Summary of work / meetings of Islamabad Campus

Quarters	Agenda	Description
4th 2018	• New Architecture discussion • Complete report on FLP-EPN (all simulations). • COMSATS Islamabad presented perfect switch FULL scale simulation	• New architecture (FLPs and EPNs are placed at different locations). • Few other changes like for example change of bandwidth between FLPs and EPNs and room like structure for EPNs. • Presented a report on complete simulation from first to last simulation results in order to show the similarity and effectiveness of results.
3rd 2018	• COMSATS Islamabad group discussion on simulation time and storage	• Successfully reduce the time of simulation by using the command line instead of GUI simulation • Result file was reduced to MBs (Initially 300-400 GBs).
1st and 2rd 2018	• COMSATS Islamabad proposed multi-layered model for FLP-EPN simulation.	• Proposed multi-layered architecture by COMSATS Islamabad • Results were also discussed. • The building of software for O2. • Simulation workflow (building time frame). • RDMA protocol (DMA transfer b/w nodes). • Omnipath of Intel (discussion on its benefits). • Comparison of Omnipath and InfiniBand

Current Standing

Task Assigned	Progress
New architecture implementation and Development (FLPs and EPNs are physically separated)	In Progress (50% completed)
Full scale architecture simulation. (Single Switch Implementation)	Completed
Multi-level hierarchical network topology along with traffic shaping.	Completed
Small-scale simulation results of the proposed model	Completed
Implement the scale up simulation using the HPC facility available at the CUI Islamabad.	Completed
Perfect switch design implementation.	Completed
R&D on simulation tools available.	Completed

Thank you