GÉANT
Networks • Services • People

In-band Network Telemetry Measurements and Summary

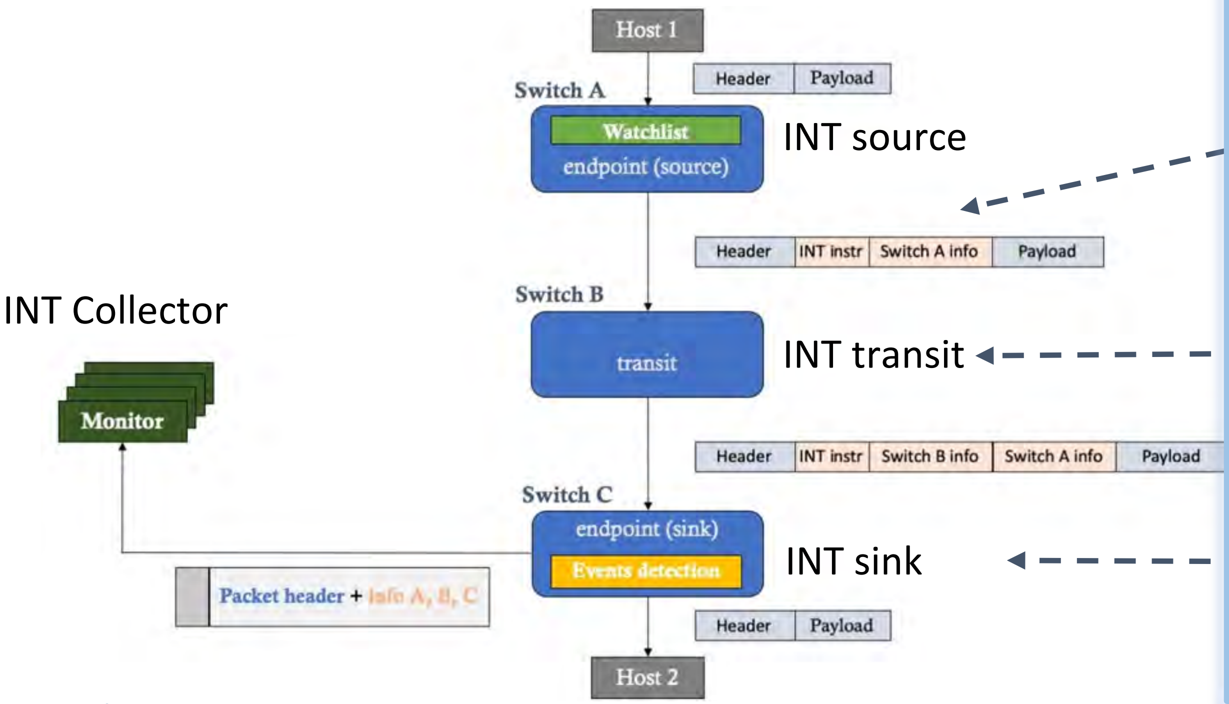
Mauro Campanella (GARR), Tomas Martinek and Mario Kuka (CESNET), Joseph Hill (UvA), Matteo Gerola (FBK), Jakub Kabat (PSNC), Marinos Demolianis and Nikos Kostopulos (NTUA), Theodore Vasilopoulos(GRNET)

GÉANT Infoshare
24 November 2022

www.geant.org

In-Band Network Telemetry (INT) short summary

INT alters structure of selected packets, in the fly, to collect and transport information in-band, in real time



INT functions

INT source node adds a small **INT header** to **every chosen packet** containing e.g. Switch IDs, Interfaces IDs, Timestamps, Link and queue utilization

INT transit nodes add specific local

The last **INT sink** node extracts, may analyze and and sends information to a collector

INT as data plane programming application

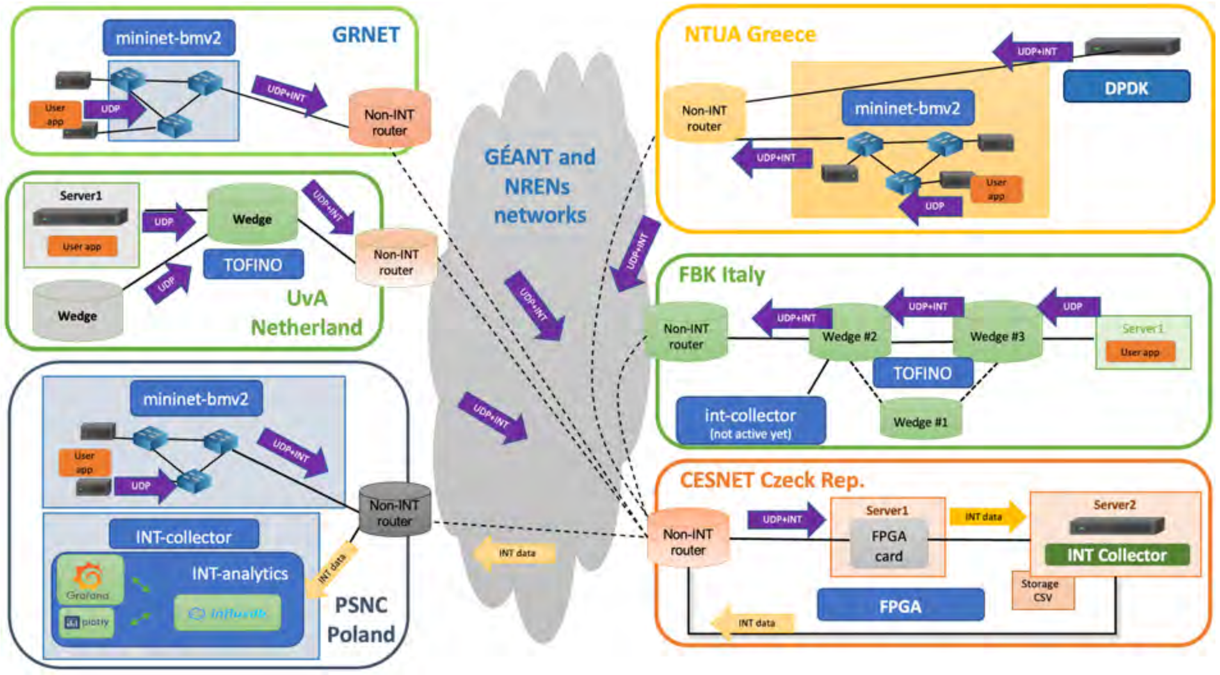
- Alters the structure of every packet with "in-house" logic
- Applies "non switching" logic to packets
- Personalizes monitoring and fast-feed information to control plane
- Offers a tool that permits precision monitoring in faster, larger and more automated networks.

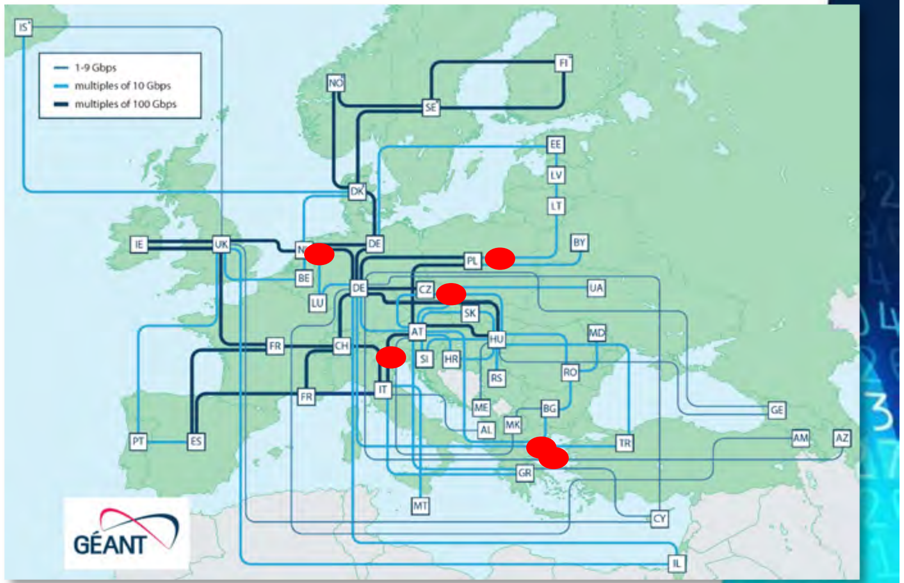
Investigation with tools started to be available in 2018:

Programming Protocol-Independent Packet Processors :
High level, C-like, coding language for controlling packet forwarding planes in networking devices

- New silicon, P4 enabled (Tofino, FPGA …), working at line speed enabling programming, Telemetry (push vs pull model),…

Our INT testbed over production NREN networks





- Packet carries timestamps, sequence number in INT headers between Source (all) and Sink node (CESNET)
- UDP packets generated at constant rate ~1k to 300k pps

- 4 switch platforms
- UDP packets flow in NRENs networks
- Collected INT data in CESNET is sent to PSNC for collection and presentation.

Measurements : Why, How

WHY: Measure at the microsecond level the effect of the transport on NRENs' networks of "user" packets

HOW: Tag with INT information each packet of an e2e UDP flow between sites in different countries.

Measure:

- the inter-packet gap (IPG) variation, computed as the difference of the distance in time between two consecutive packet at the source and at the destination
- Packet losses and reordering
- End to end packet delay variation (IPDV RFC 3393)

INT headers content and positioning

Header	number of bytes
Ethernet	18
IP	20
UDP	8
INT	12
INT Node Metadata	n x 24
INT tail	4
User DATA (left empty)	

Follows INT standard 1.0 **extended**

+ timestamps of 64 bits

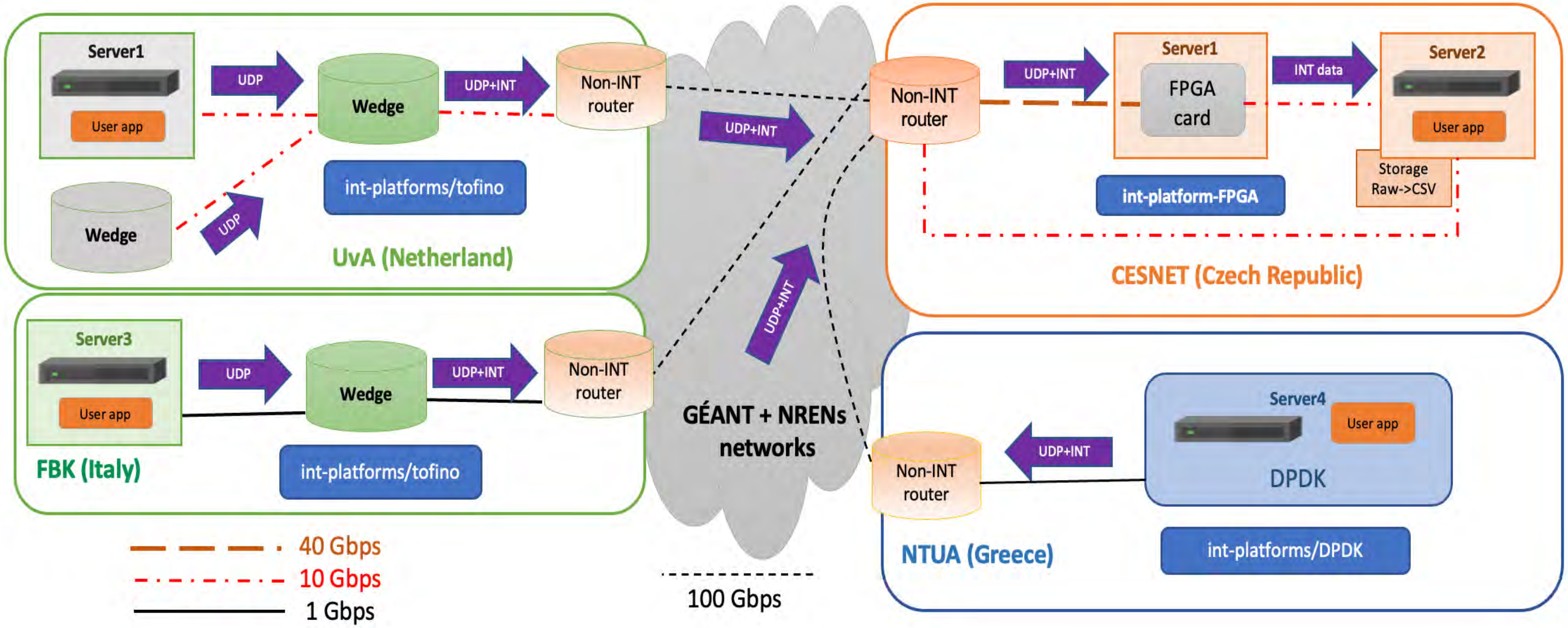
+ source sequence number (16bits)

INT Headers inserted between UDP and user data to avoid any modification of the standard headers. Containing:

switch id, ingress and egress port id, source and destination sequence numbers and timestamps

A single UDP flow packets are sent (using iperf2 or a packet generator) to a switch that inserts the source INT headers, are “normally” switched to the destination sink node in CESNET (an FPGA with 300 GB of RAM)

INT measurements topology



FBK, NTUA and UvA as INT sources, CESNET as INT Sink

INT P4 code developed for these platforms (on GitHuB)

P4 on DPDK
"classic" HW

DPDK (Data Plane Development Kit) kernel software acceleration, available for the P4 to DPDK compiler (T4P4S)

BMv2 virtual
P4 software switch

BMv2 Behavioral Model v2 – emulation of Tofino
Uses Mininet

Netcope NFB-100G2

FPGA P4 compiler developed by CESNET up to 2x100 Gbps

Edgecore Wedge100BF-32X

Tofino ASIC family engineered for dataplane programming by Barefoot (now Intel)

Methodology

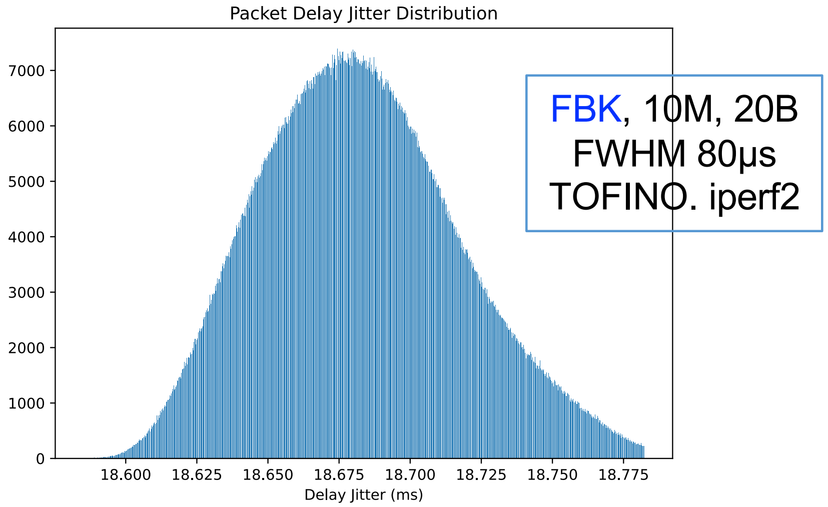
- UDP flow
- 2 capacities tested: 10Mbps and 100 Mbps, 120s each
- For each capacity : 2 packet sizes of 20 and 1000 bytes (user payload)
- Where possible (UvA) increase testing in capacity at Gbps and optionally run 1 test of 4 hours at 1 Mbps at 1000 bytes
- Use the FPGA also as a INT data collector.
  Choice due to scalability limitation of the single application collector in PSNC (need to use a distributed event streaming platform like KAFKA) and to limit the packet path after INT sink tagging.

Losses and reordering

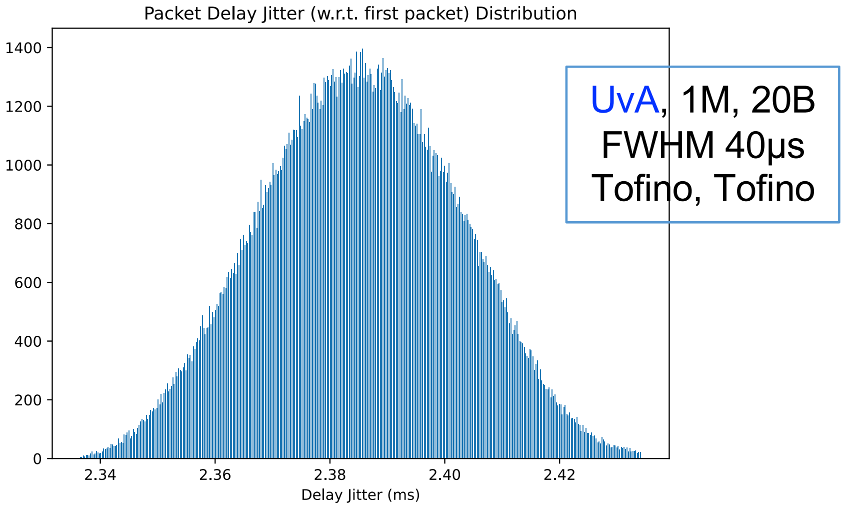
In all measurements there are no lossess and almost no reordering (less than 1 case per million packets).

There has been a congestion issue on the UvA LAN which did not allow to perform stable tests at capacity higher than 100 Mbps.

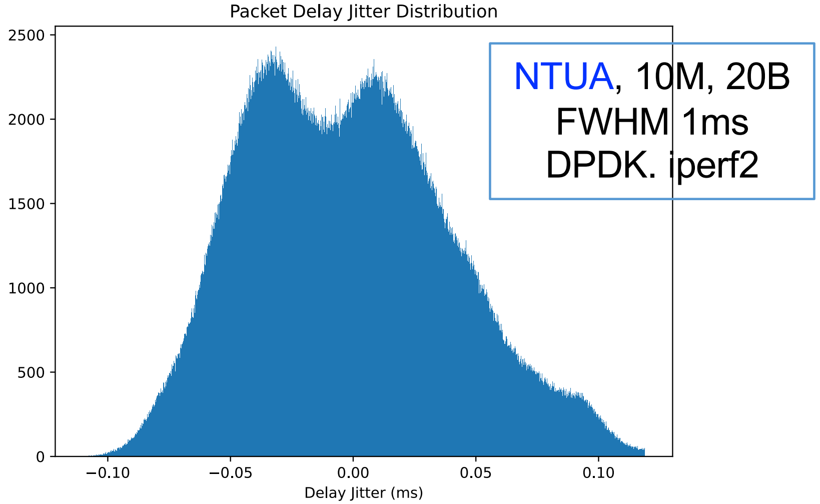
IP Packet delay variation - absolute delay values not accurate



FBK, 10M, 20B
FWHM 80µs
TOFINO. iperf2



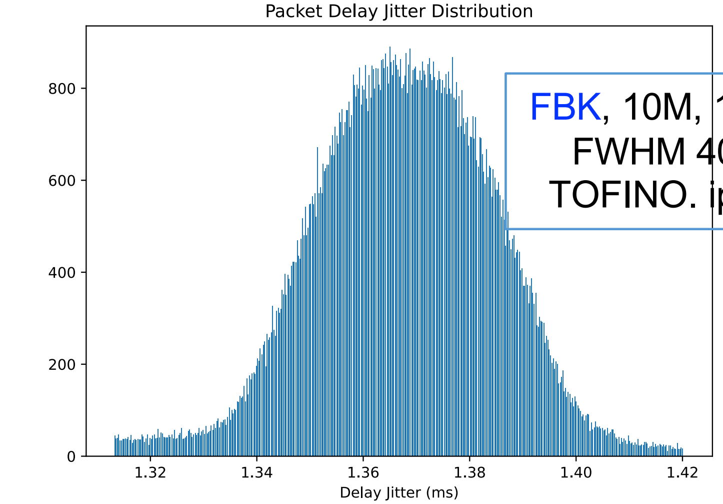
UvA, 1M, 20B
FWHM 40µs
Tofino, Tofino



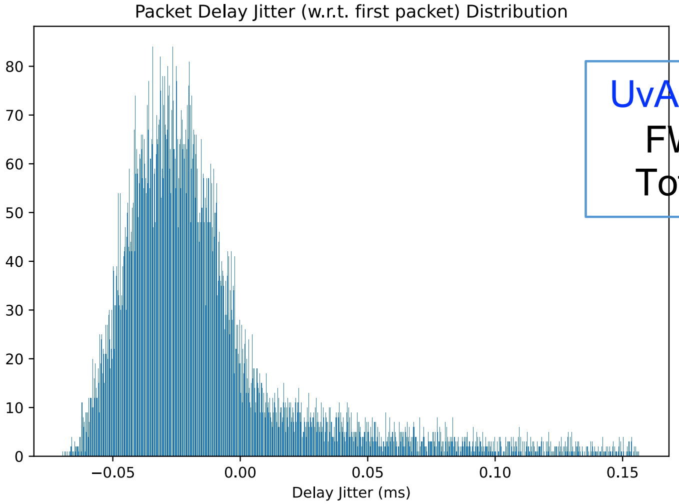
NTUA, 10M, 20B
FWHM 1ms
DPDK. iperf2

- Excellent delay variation on long paths
- Sofware packet generation increases jitter

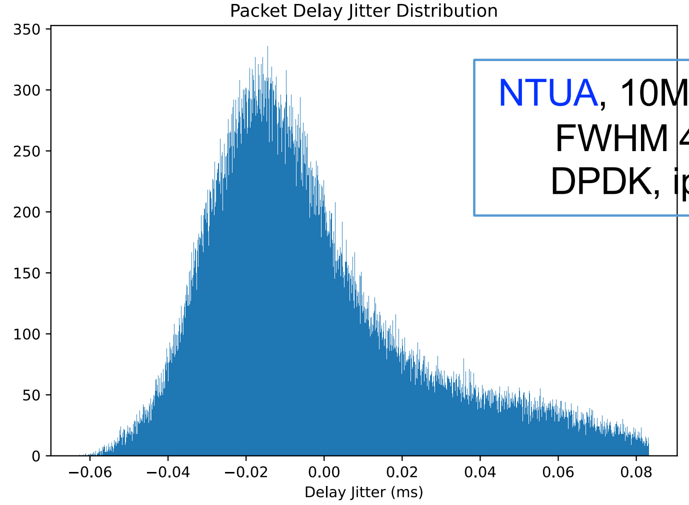
IP Packet delay variation - absolute delay values not accurate



FBK, 10M, 1000B
FWHM 40µs
TOFINO. iperf2



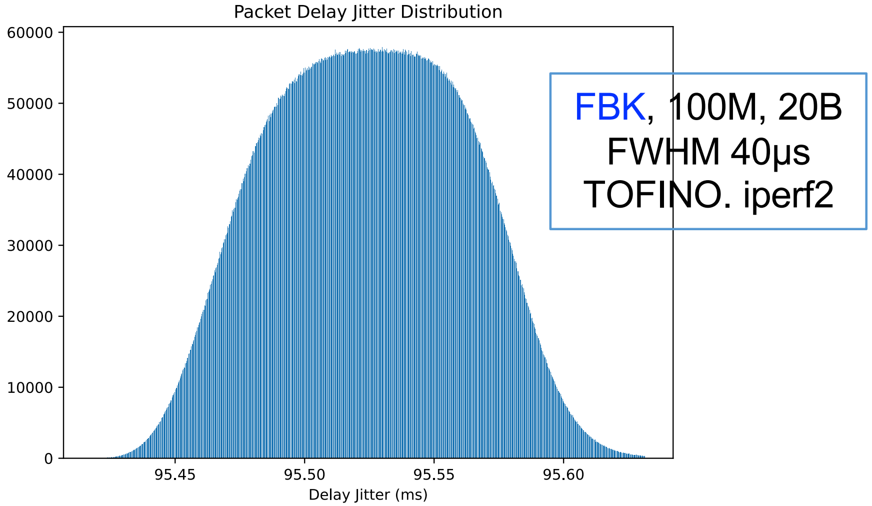
UvA, 1M, 100OB
FWHM 40µs
Tofino, Tofino



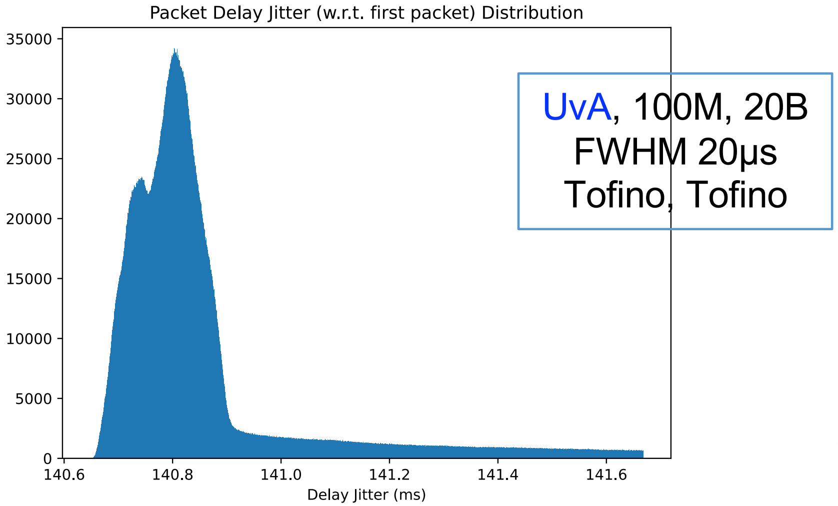
NTUA, 10M, 1000B
FWHM 40µs
DPDK, iperf2

- Excellent delay variation on long paths also for large size packets
- Long tails

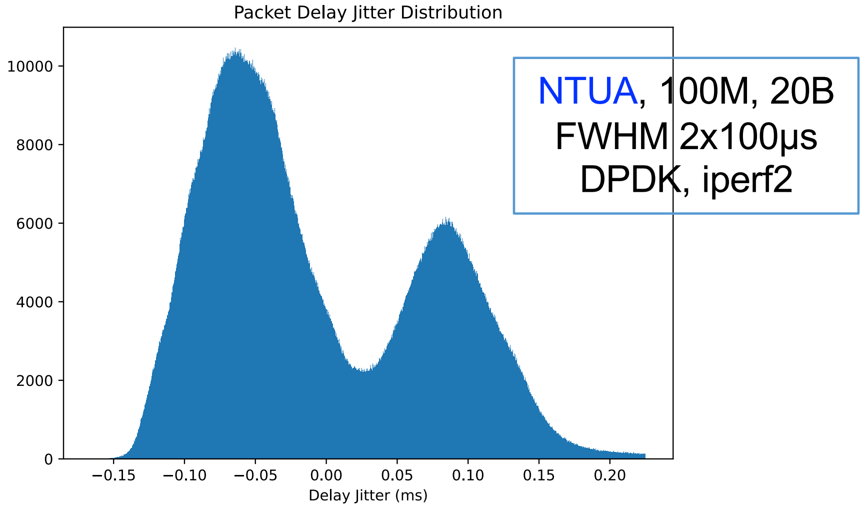
IP Packet delay variation - absolute delay values not accurate



FBK, 100M, 20B
FWHM 40µs
TOFINO. iperf2



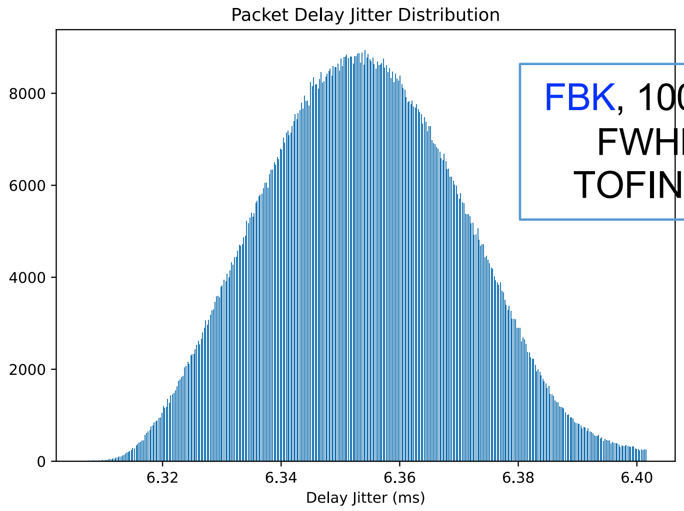
UvA, 100M, 20B
FWHM 20µs
Tofino, Tofino



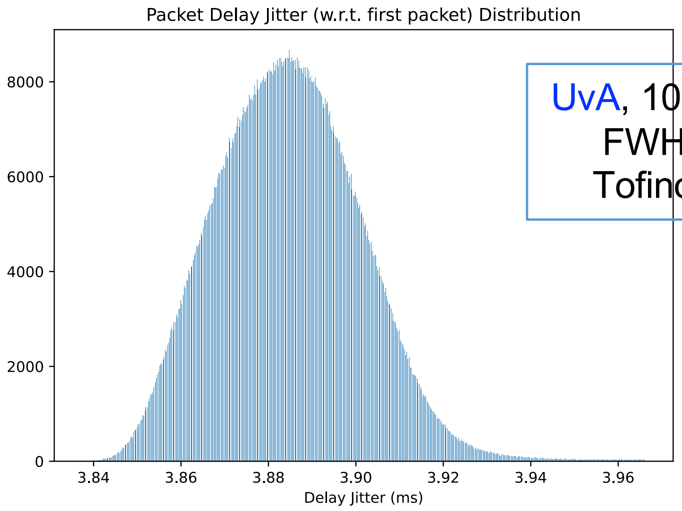
NTUA, 100M, 20B
FWHM 2x100µs
DPDK, iperf2

- Increasing capacity maintains jitter small
- Small packet exhibit a worse behavior n th software platform (still adequate)

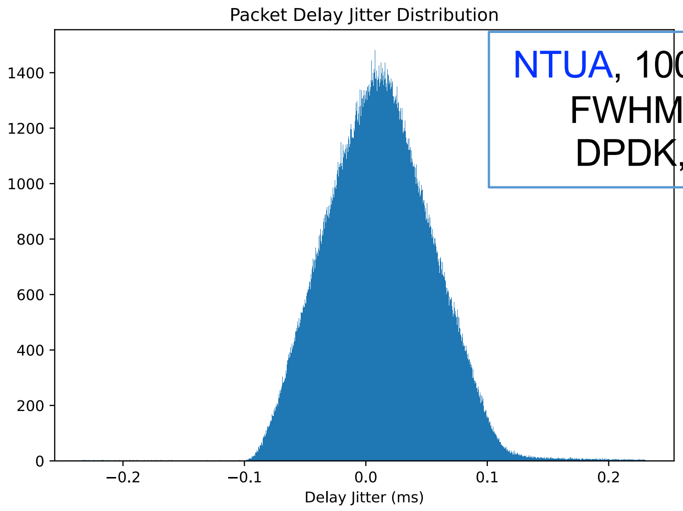
IP Packet delay variation - absolute delay values not accurate



FBK, 100M, 1000B
FWHM 35µs
TOFINO. iperf2



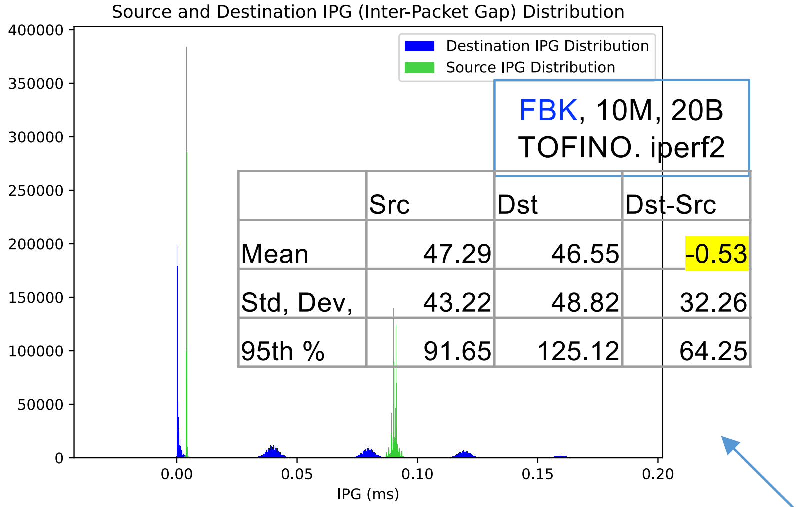
UvA, 100M, 1000B
FWHM 35µs
Tofino, Tofino



NTUA, 100M, 1000B
FWHM 100µs
DPDK, iperf2

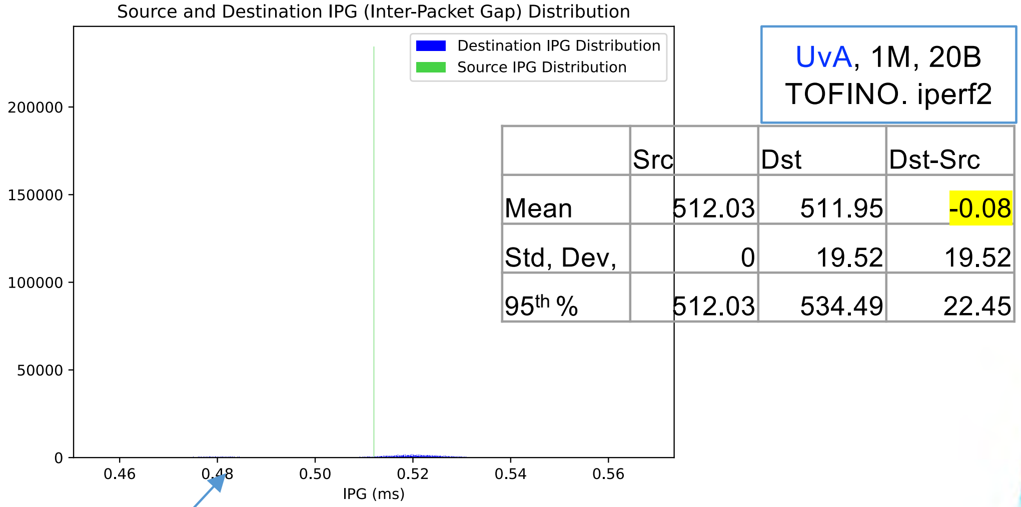
- Higher capacity and packet sizes provides an excellent Jitter

Inter packet Gap distributions at source and sink



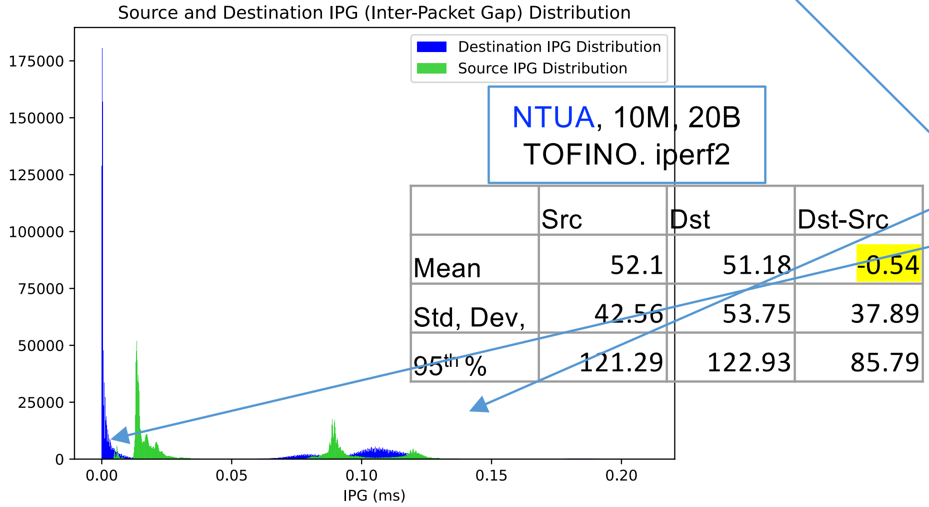
FBK, 10M, 20B
TOFINO. iperf2

	Src	Dst	Dst-Src
Mean	47.29	46.55	-0.53
Std, Dev,	43.22	48.82	32.26
95th %	91.65	125.12	64.25



UvA, 1M, 20B
TOFINO. iperf2

	Src	Dst	Dst-Src
Mean	512.03	511.95	-0.08
Std, Dev,	0	19.52	19.52
95th %	512.03	534.49	22.45

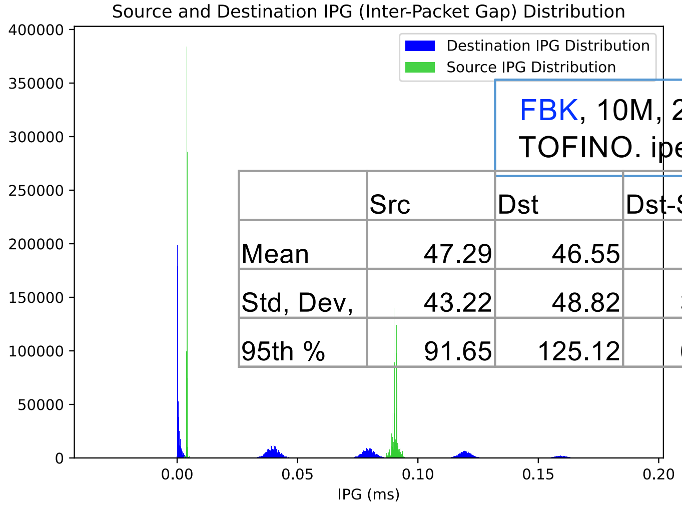


NTUA, 10M, 20B
TOFINO. iperf2

	Src	Dst	Dst-Src
Mean	52.1	51.18	-0.54
Std, Dev,	42.56	53.75	37.89
95th %	121.29	122.93	85.79

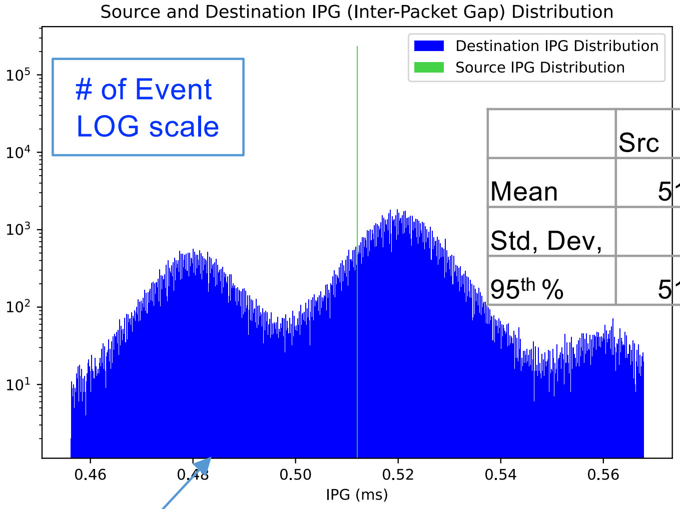
- Packet generation in HW (Tofino) provides very stable and excellent source IPG
- Software generators shows issues
- The network creates back-to-back packets
- The average value of IPG is maintained after the transport on NRENs and local LANs (their difference is highlighted in yellow)

Inter packet Gap distributions at source and sink



FBK, 10M, 20B
TOFINO. iperf2

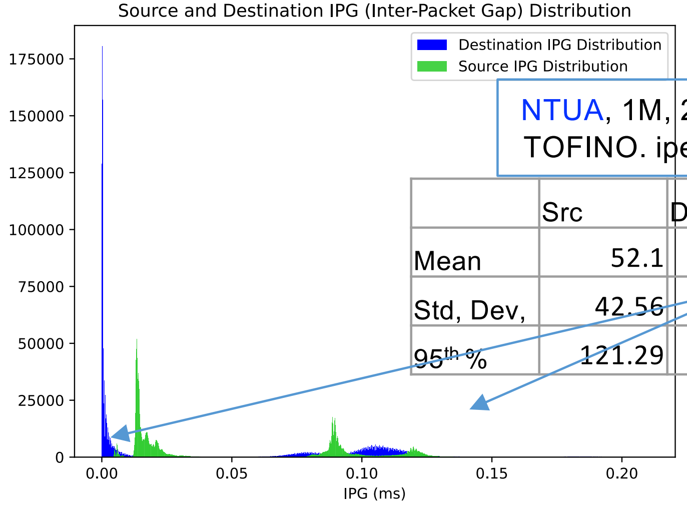
	Src	Dst	Dst-Src
Mean	47.29	46.55	-0.53
Std, Dev,	43.22	48.82	32.26
95th %	91.65	125.12	64.25



# of Event
LOG scale

UvA, 1M, 20B
TOFINO. iperf2

	Src	Dst	Dst-Src
Mean	512.03	511.95	-0.08
Std, Dev,	0	19.52	19.52
95th %	512.03	534.49	22.45

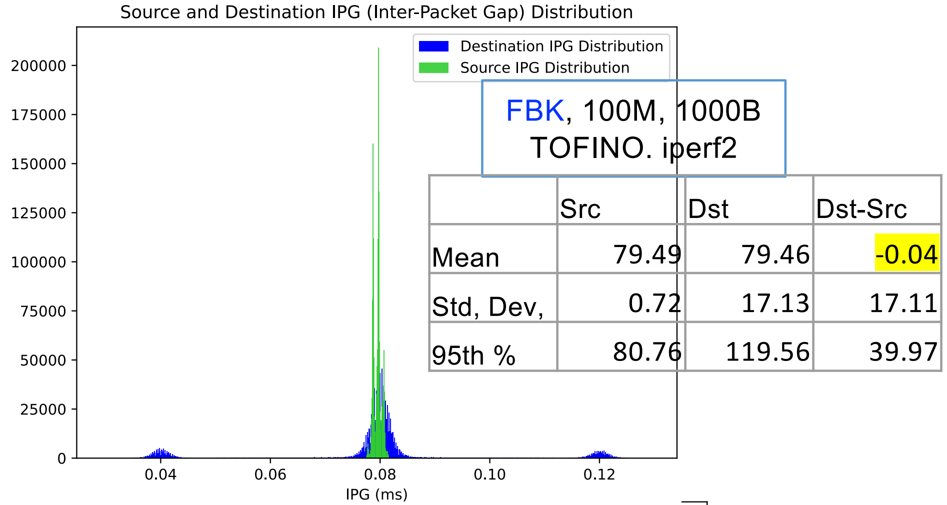


NTUA, 1M, 20B
TOFINO. iperf2

	Src	Dst	Dst-Src
Mean	52.1	51.18	-0.54
Std, Dev,	42.56	53.75	37.89
95th %	121.29	122.93	85.79

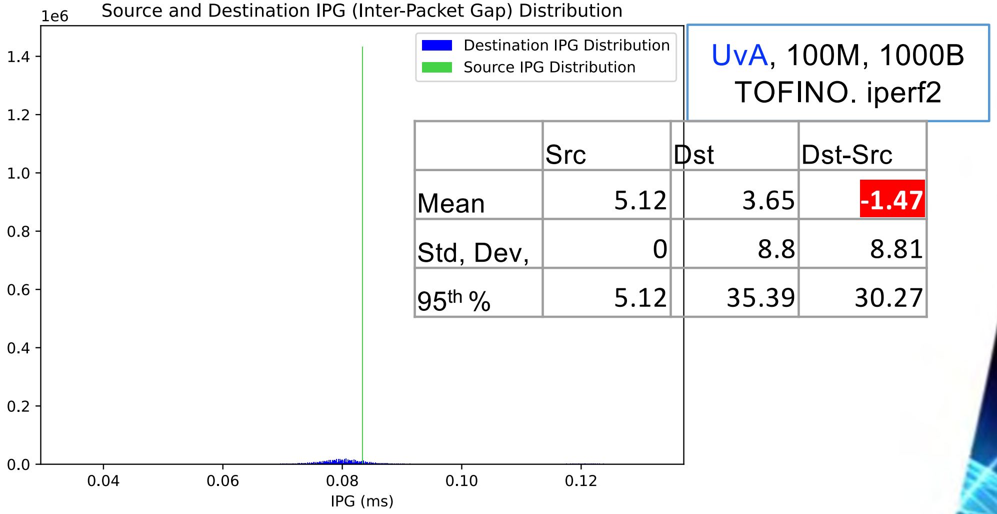
- Packet generation in HW (Tofino) provides very stable and excellent source IPG
- Software generators shows issues
- The network creates back-to-back packets
- The average value of IPG is maintained after the transport on NRENs and local LANs (their difference is highlighted in yellow)

Inter packet Gap distributions at source and sink



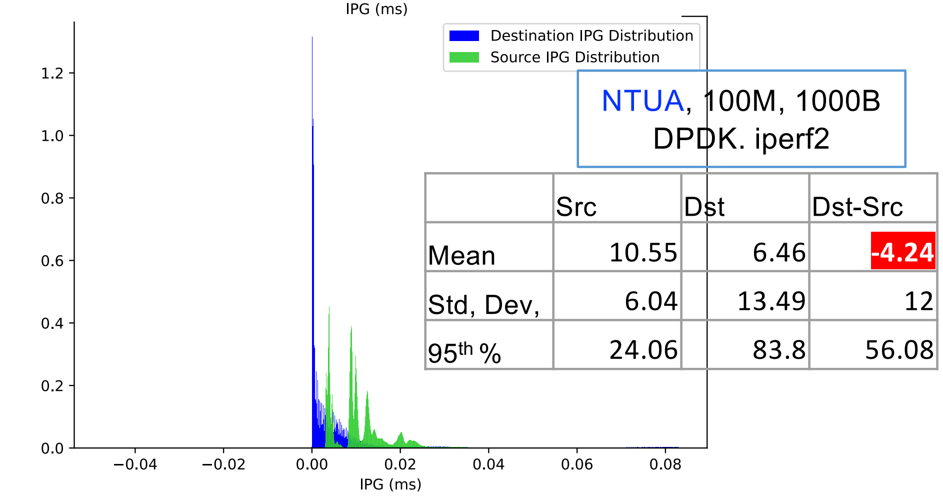
FBK, 100M, 1000B TOFINO. iperf2

	Src	Dst	Dst-Src
Mean	79.49	79.46	-0.04
Std, Dev,	0.72	17.13	17.11
95th %	80.76	119.56	39.97



UvA, 100M, 1000B TOFINO. iperf2

	Src	Dst	Dst-Src
Mean	5.12	3.65	-1.47
Std, Dev,	0	8.8	8.81
95th %	5.12	35.39	30.27

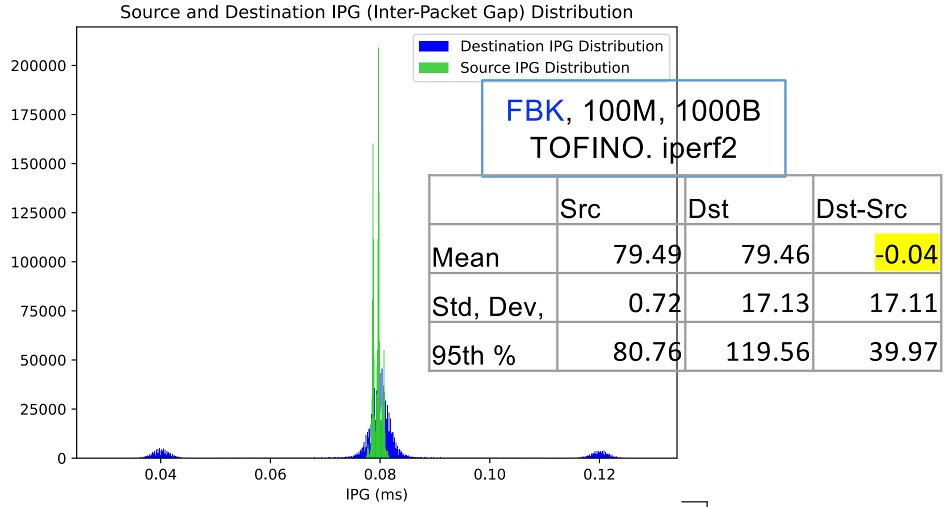


NTUA, 100M, 1000B DPDK. iperf2

	Src	Dst	Dst-Src
Mean	10.55	6.46	-4.24
Std, Dev,	6.04	13.49	12
95th %	24.06	83.8	56.08

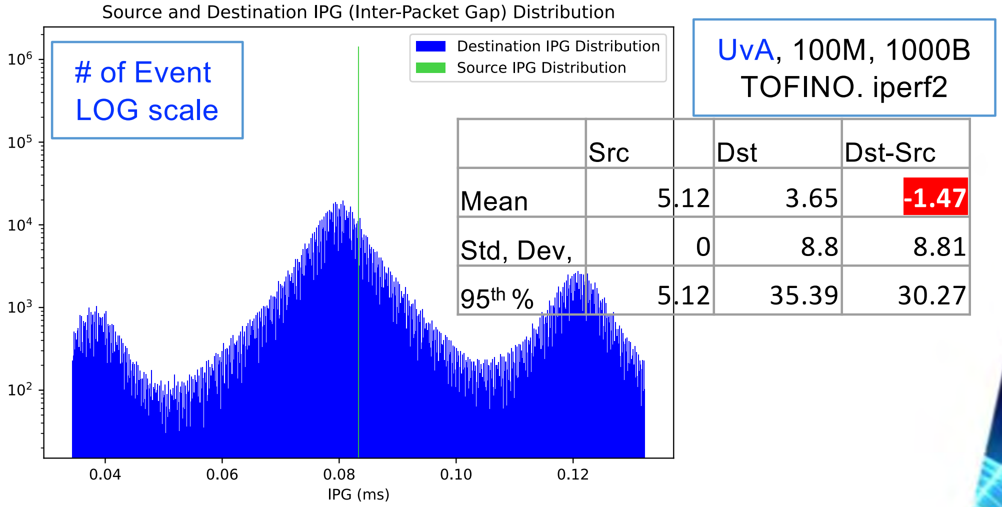
- Packet generation in HW (Tofino) provides very stable and excellent source IPG, independent of packet size
- Software generators shows issues at all sizes
- The network creates back-to-back packets
- The average value of IPG is lower at destination due to a network effect (marked in red)

Inter packet Gap distributions at source and sink



FBK, 100M, 1000B
TOFINO. iperf2

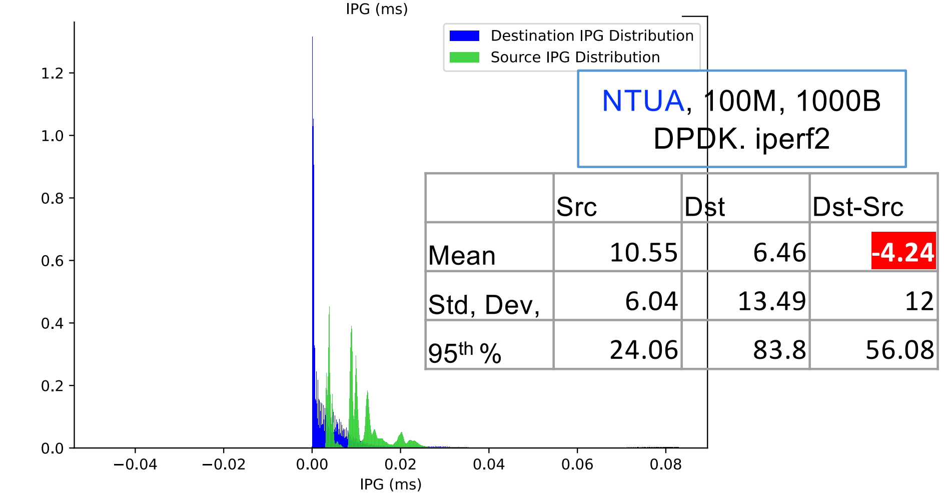
	Src	Dst	Dst-Src
Mean	79.49	79.46	-0.04
Std, Dev,	0.72	17.13	17.11
95th %	80.76	119.56	39.97



# of Event
LOG scale

UvA, 100M, 1000B
TOFINO. iperf2

	Src	Dst	Dst-Src
Mean	5.12	3.65	-1.47
Std, Dev,	0	8.8	8.81
95th %	5.12	35.39	30.27



NTUA, 100M, 1000B
DPDK. iperf2

	Src	Dst	Dst-Src
Mean	10.55	6.46	-4.24
Std, Dev,	6.04	13.49	12
95th %	24.06	83.8	56.08

- Packet generation in HW (Tofino) provides very stable and excellent source IPG, independent of packet size
- Software generators shows issues at all sizes
- The network creates back-to-back packets
- The average value of IPG is lower at destination due to a network effect (**marked in red**)

Summary

- These measurements of the transport of packets on the NRENs networks show they preserve very well the packet timing profile (spreading packet delay by few tens of microseconds between source and destination)
- No packet loss and reordering has been recorded
- The cumulative network switching behaviour at microsecond, or lower, scale exhibits a complex behaviour. Packets may be groomed or delayed, however the flow maintains its timing profile. These effects are in the tenths of microsecond range and may affect only real time communication (interactive applications or control plane signalling)
- INT can be used also by users between cooperating end-sites without modifying production transport networks.

Summary (cont.)

- Packet handling in software works, with some effects, up to few Gbps.
- ASICs are required above few Gbps and to ensure the best timing precision.
- INT/P4 use may/will generate and require handling of large amount of "raw" data, to be used for analytics and more.
- Event at “low capacities” a single collector application has scalability issues, the back-end component is essential (e.g. use of KAFKA, see also the ESnet experience)

Acknowledgements

The effort reported has received essential contributions from many GÉANT participants in the last 4 years and specifically:

Damu Ding (Italy), Federico Pederzolli (Italy), Pavel Benacek (Czech Republic), Marco Savi (Italy)

Tim Chown (Jisc UK), Ivana Golub (PSNC Poland),

Xavier Jeannin (RENATER France)

A sicere Thank You !

More information on GÉANT Data Plane Programmability activity

- **Data Plane Programming** / **INT GEANT web page** https://wiki.geant.org/display/NETDEV/INT Includes all documents produced and a **pointer to GitHub INT P4 code**
- **Mailing list:** https://lists.geant.org/sympa/subscribe/int-discuss,
- **White Paper INT Tests in NREN networks** – DPP WP6 T1 white paper https://www.geant.org/Resources/Documents/GN4-3_White-Paper_In-Band-Network-Telemetry.pdf
- **DDoS Paper**: "In-Network Volumetric DDoS Victim Identification Using Programmable Commodity Switches", F. Pederzolli, M. Campanella and D. Siracusa, in IEEE Transactions on Network and Service Management, Vol. 18, Issue: 2, June 2021,
page: 1191-1202, DOI: 10.1109/TNSM.2021.3073597 and at https://arxiv.org/abs/2104.06277

More information on GÉANT Data Plane Programmability activity

- **White Paper: Timestamping and Clock Synchronisation in P4-Programmable Platforms**– DPP WP6 T1 white paper – 8 September 2022
  https://resources.geant.org/wp-content/uploads/2022/09/GN4-3_White-Paper_Timestamping-and-Clock-Synchronisation-in-P4-Programmable-Platforms.pdf
- The GÉANT **First Telemetry and Big Data Workshop**
  https://wiki.geant.org/display/PUB/Telemetry+and+Big+Data+Workshop
- The GÉANT **2nd Telemetry and Data Workshop – 6 April 2022**
  https://events.geant.org/event/1104/

Non GEANT References

- **The Programmable Data Plane Reading List** : https://programmabledataplane.review/
- Oliver Michel, Roberto Bifulco, Gábor Rétvári, Stefan Schmid, **"The Programmable Data Plane: Abstractions, Architectures, Algorithms, and Applications"**, ACM Computing Surveys, Volume 54, Issue 4, May 2021, Article No.: 82, pp 1–36, https://doi.org/10.1145/3447868 10.36227/techrxiv.12894677.v1 https://www.univie.ac.at/ct/stefan/csur21.pdf
- **"A Survey on Data Plane Programming with P4: Fundamentals, Advances, and Applied Research"**, Frederik Hauser, Marco Häberle, Daniel Merling, Steffen Lindner, Vladimir Gurevich, Florian Zeiger, Reinhard Frank, and Michael Menth (50 pages).26 Jan 2021, to be published in" Communications Surveys & Tutorials (COMST) journal --https://arxiv.org/pdf/2101.10632.pdf
- Kaur, Sukhveer & Saluja, Krishan & Aggarwal, Naveen. (2021). **"A review on P4-Programmable data planes: Architecture, research efforts, and future directions"**. Computer Communications. 170. 10.1016/j.comcom.2021.01.027.

Thank you

Questions ?

GÉANT
Networks • Services • People

www.geant.org

© GÉANT Association on behalf of the GN4 Phase 3 project (GN4-3).
The research leading to these results has received funding from
the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 856726 (GN4-3).