

10 Gbit Line Rate Packet-to-Disk Using n2disk

Background

Storing the entire raw traffic is required for:

- Lawful interception and network forensics.
- New threat discovery.
- Complex network troubleshooting.

Motivations [1/2]

- A 10 Gigabit link carries over 1 GB/sec.
- Dropping packets is not allowed as it will make the recorded traces mostly useless
- Packet recorders must operate at line-rate with any packet size and traffic conditions.

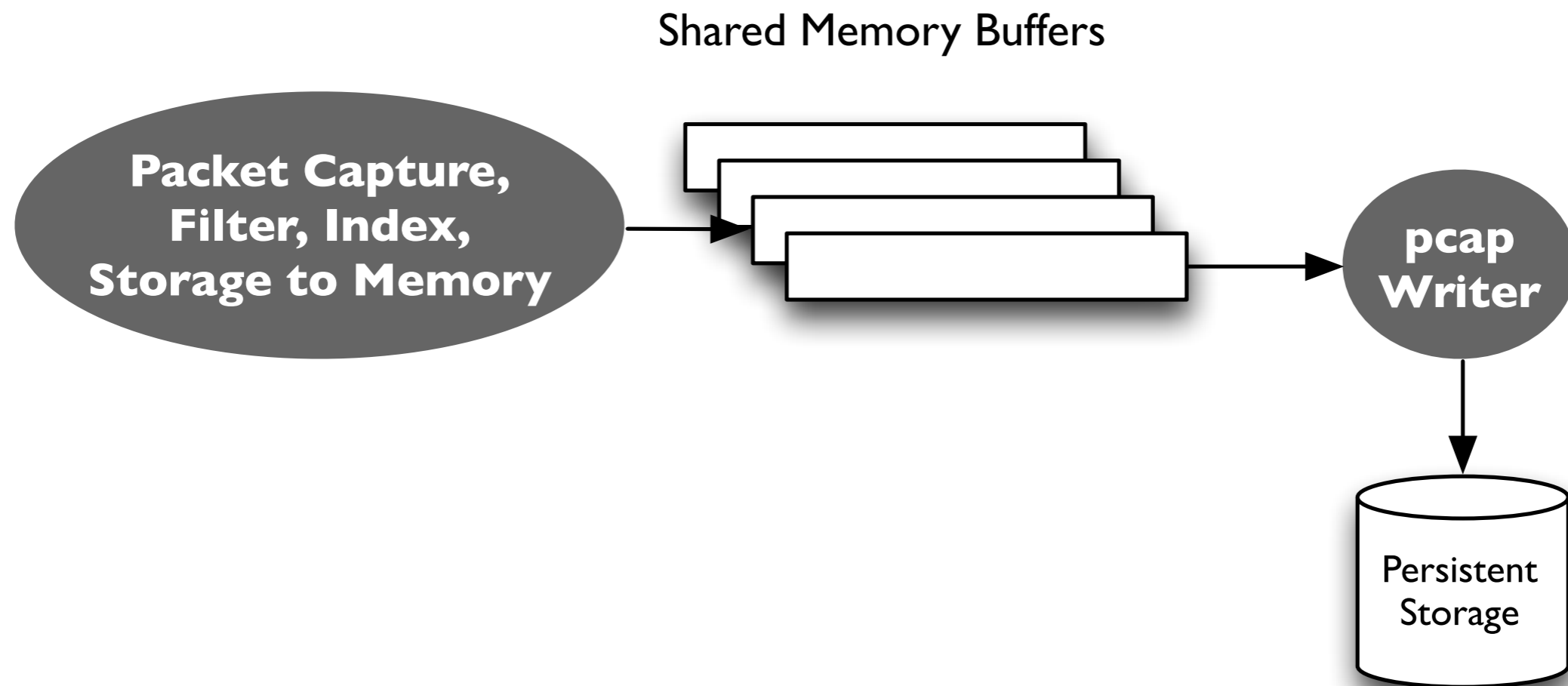
Motivations [2/2]

- If you store big data you want to access/retrieve packet quickly.
- Searching packets in pcap corresponds to a linear file scan: we need indexes.
- Many available solutions use proprietary dump formats: want to use an open format, a.k.a. pcap.

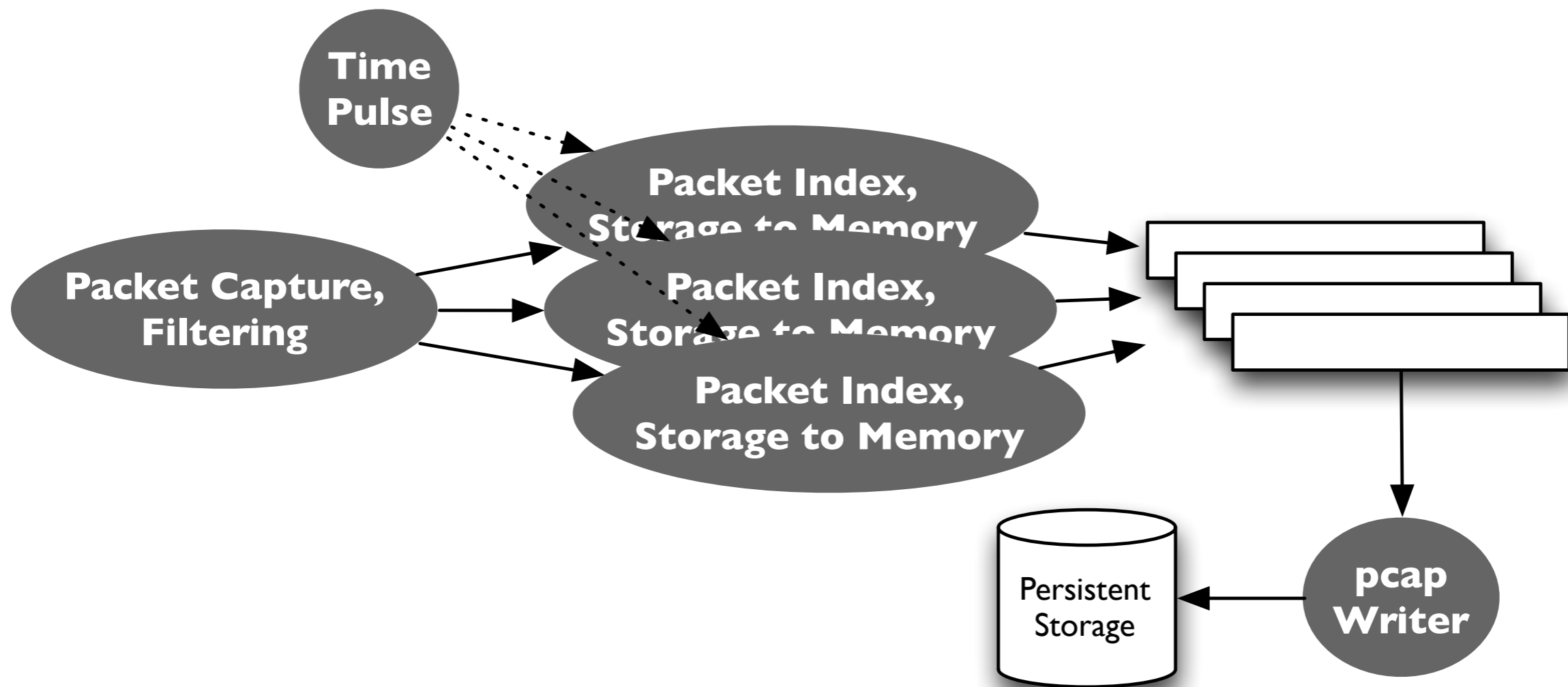
n2disk

- Open .pcap file format.
- On-the-fly probabilistic indexing (digest).
- Nanosecond hw timestamps on Intel 1Gbit 82580/i350 or Silicom 10Gbit HW TS NIC.
- Based on DNA/libzero packet capture technology for 100% packet capture.

Single-Threaded Mode

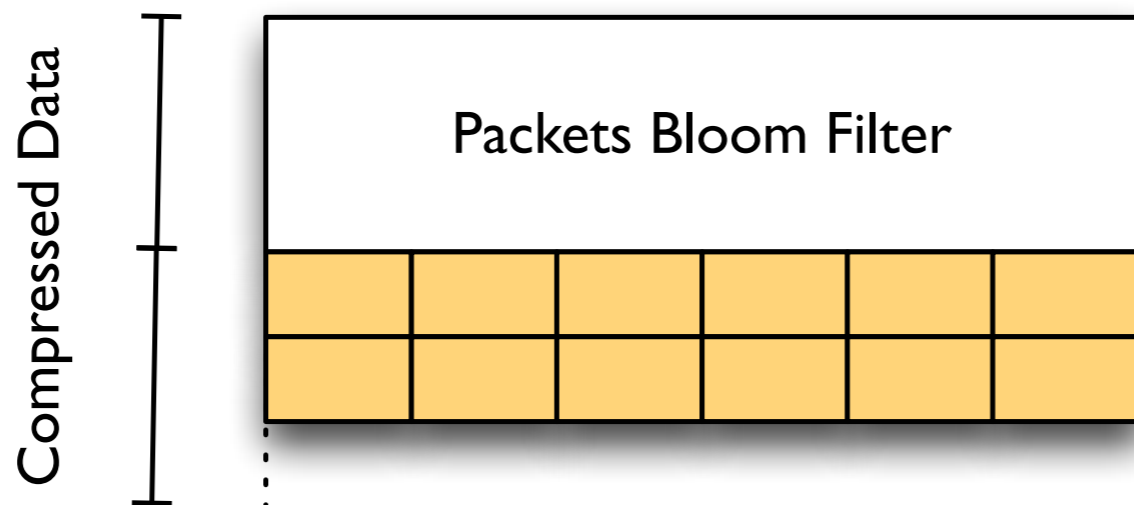
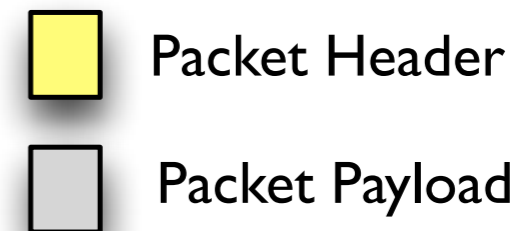
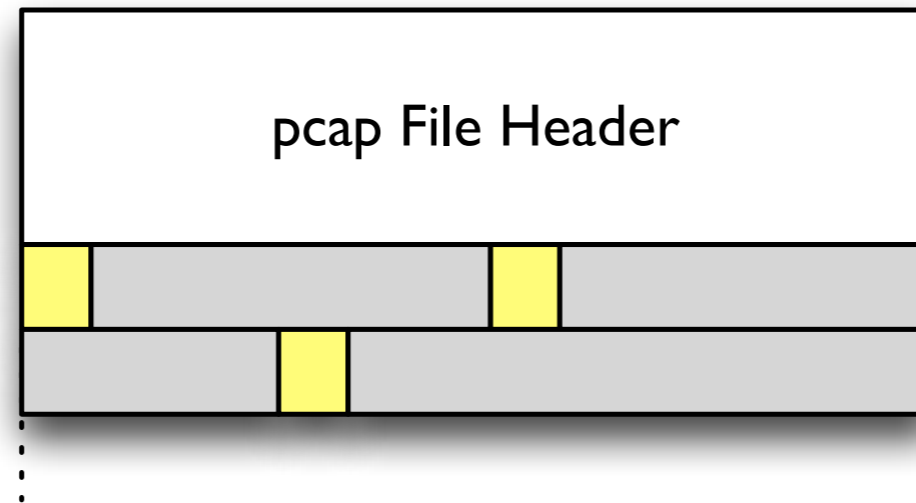
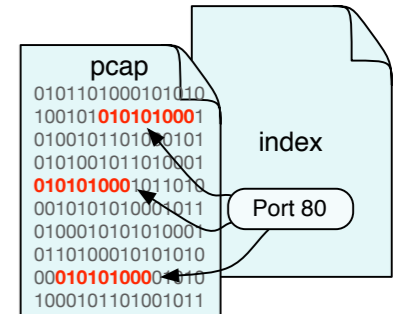


Multi-Threaded Mode



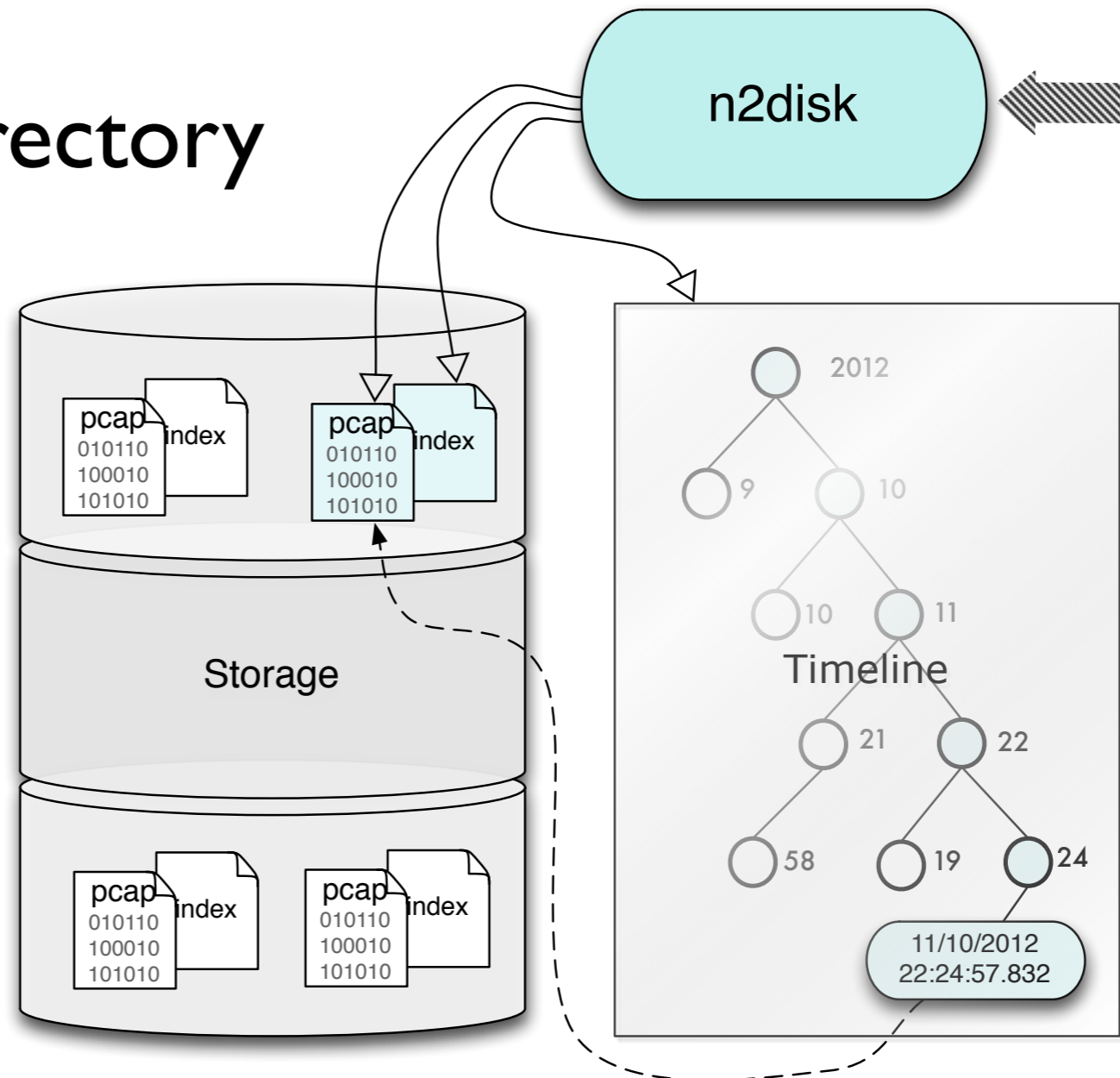
Index

Every pcap file comes with an optional companion index file



Timeline

A time-ordered directory tree maintained by n2disk to enable time-based packet extraction



nBox

- The nBox is a open-source companion web GUI that we have created for configuring ntop apps and drivers.
- We have added support for n2disk thus users can play with it instead of using the command line interface.



nBox: Filter Wizard

Index Filter Wizard x

IPv4	SrcOrDst	192.168.1.23	/32
IPv6	SrcOrDst		/128
MAC	SrcOrDst		Port SrcOrDst 80 NOT

AND **OR** **BLOCK START** **BLOCK END**

IPv4	SrcOrDst	192.168.1.23	/32	AND
Port	SrcOrDst	80		

nBox: Packet Extraction

Apps / n2disk / Extract (eth0)

Extract Packets

From

To

Filter

Output File

Specify where the file will be created.

used by the popular [tcpdump](#) tool).

with does not exist, it will be

January 2013						
Su	Mo	Tu	We	Th	Fr	Sa
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31	1	2
3	4	5	6	7	8	9

NOTE: you can download the file via FTP or SSH. Please configure a login name and password in the Users Configuration web page.

Start Extraction

nBox: Packet Browsing

Apps / n2disk / Dump (eth0)

/storage/n2disk/eth0

Filter files by Creation Time

From 2013-01-30 23:46 To 2013-01-30 23:46 Search

File Name	Creation Time	Size
1		3.99 MB
archive		0 bytes
fake10.pcap	Wed Dec 5 16:49:08 2012	0 bytes
fake11.pcap	Wed Dec 5 16:49:08 2012	0 bytes
fake12.pcap	Wed Dec 5 16:49:08 2012	0 bytes
fake13.pcap	Wed Dec 5 16:49:08 2012	0 bytes
fake3.pcap	Wed Dec 5 16:49:08 2012	0 bytes
fake4.pcap	Wed Dec 5 16:49:08 2012	0 bytes
fake5.pcap	Wed Dec 5 16:49:08 2012	0 bytes
fake6.pcap	Wed Dec 5 16:49:08 2012	0 bytes
fake7.pcap	Wed Dec 5 16:49:08 2012	0 bytes
fake8.pcap	Wed Dec 5 16:49:08 2012	0 bytes
fake9.pcap	Wed Dec 5 16:49:08 2012	0 bytes
fakedir0		132.21 KB
fakedir1		0 bytes

imp (eth0) / Pcap Reader (/storage/n2disk/eth0/fakedir0/notsofake.pcap)

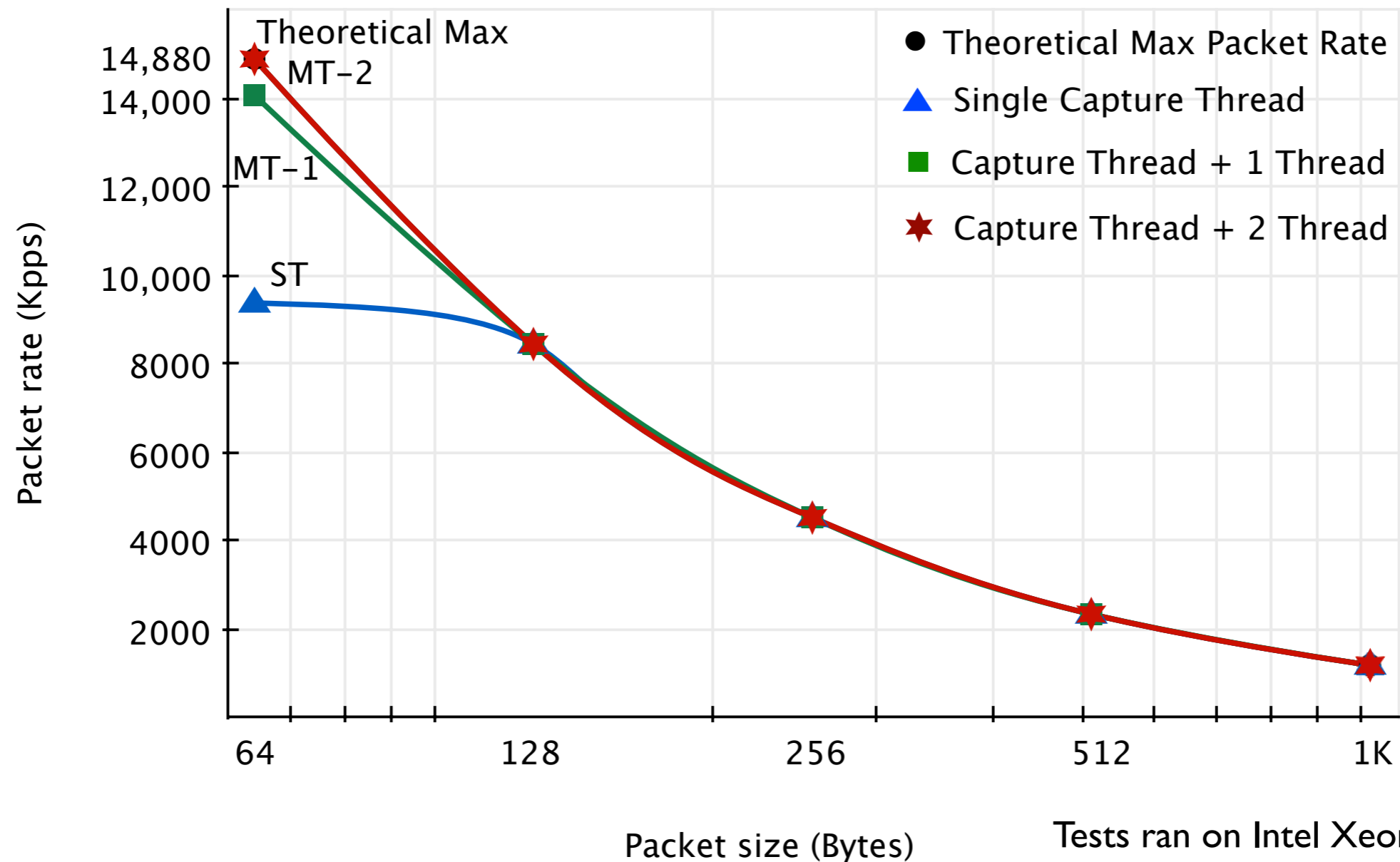
Page 1 of 20 Limit View 1 - 10 of 191

- 0.000000 192.168.0.200 -> 64.243.24.160 TCP 74 50140 > http [SYN] Seq=0 Win=65535 Len=0 MSS=1460 WS=1 TSval=1038106006 TSecr=0
- 0.171898 64.243.24.160 -> 192.168.0.200 TCP 74 http > 50140 [SYN, ACK] Seq=0 Ack=1 Win=32768 Len=0 MSS=1452 WS=1 TSval=1367466727 TSecr=1038106006
- 0.172084 192.168.0.200 -> 64.243.24.160 TCP 66 50140 > http [ACK] Seq=1 Ack=1 Win=65535 Len=0 TSval=1038106006 TSecr=1367466727
- 0.173438 192.168.0.200 -> 64.243.24.160 HTTP 361 GET / HTTP/1.1
- 0.374135 64.243.24.160 -> 192.168.0.200 TCP 66 http > 50140 [ACK] Seq=1 Ack=296 Win=32473 Len=0 TSval=1367466977 TSecr=1038106006
- 0.482397 64.243.24.160 -> 192.168.0.200 TCP 273 [TCP segment of a reassembled PDU]
- 0.493096 192.168.0.200 -> 64.243.24.160 TCP 66 50140 > http [ACK] Seq=296 Ack=208 Win=65535 Len=0 TSval=1038106007 TSecr=1367467090
- 0.512575 64.243.24.160 -> 192.168.0.200 TCP 72 [TCP segment of a reassembled PDU]
- 0.531760 64.243.24.160 -> 192.168.0.200 TCP 1506 [TCP segment of a reassembled PDU]
- 0.553891 192.168.0.200 -> 64.243.24.160 TCP 74 50141 > http [SYN] Seq=0 Win=65535 Len=0 MSS=1460 WS=1 TSval=1038106007 TSecr=0
- 0.649174 64.243.24.160 -> 192.168.0.200 TCP 1506 [TCP segment of a reassembled PDU]

Page 1 of 20 Limit View 1 - 10 of 191

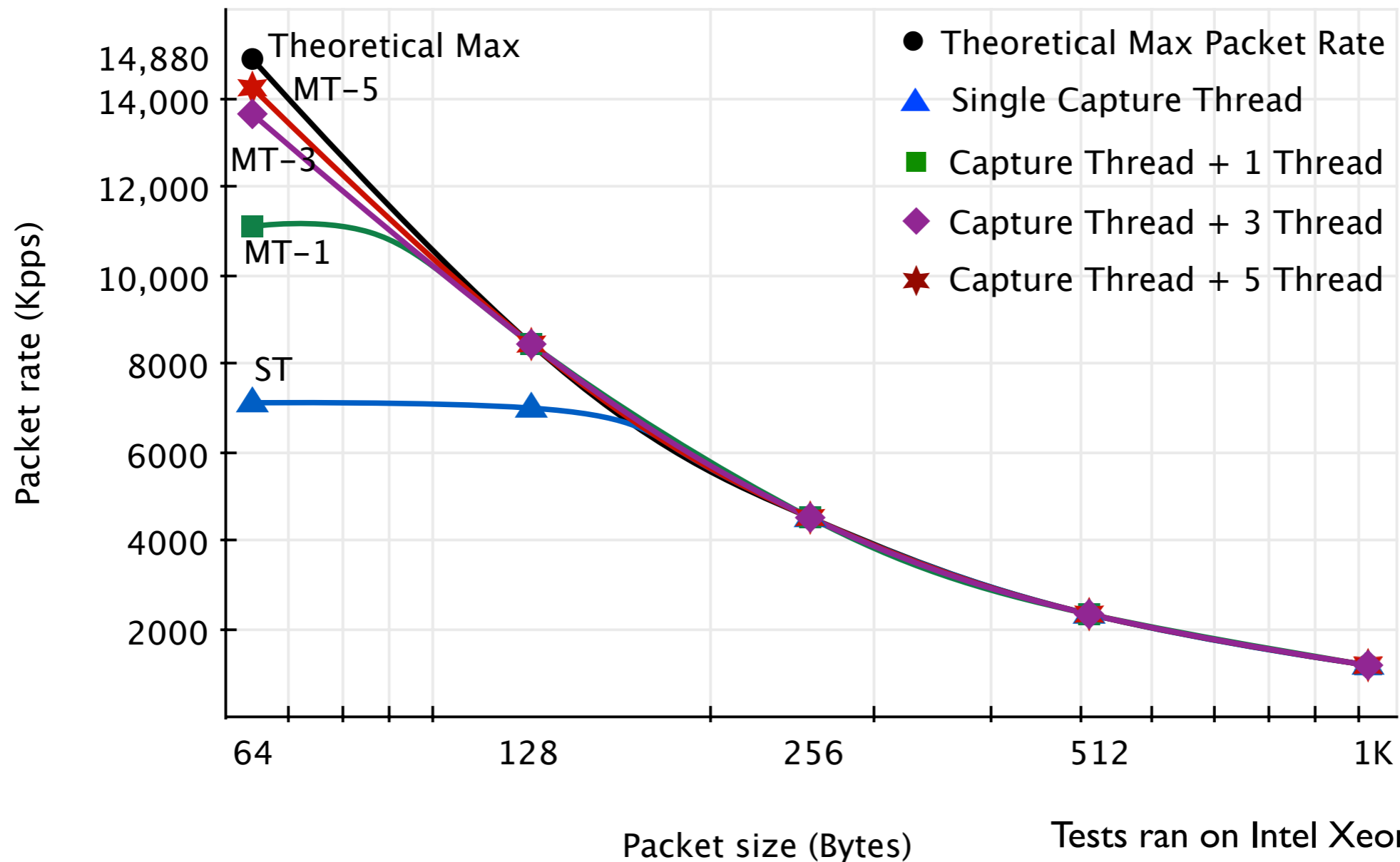
Dump Scalability

A multithreaded design is a key factor for granting line-rate on low-frequency systems.



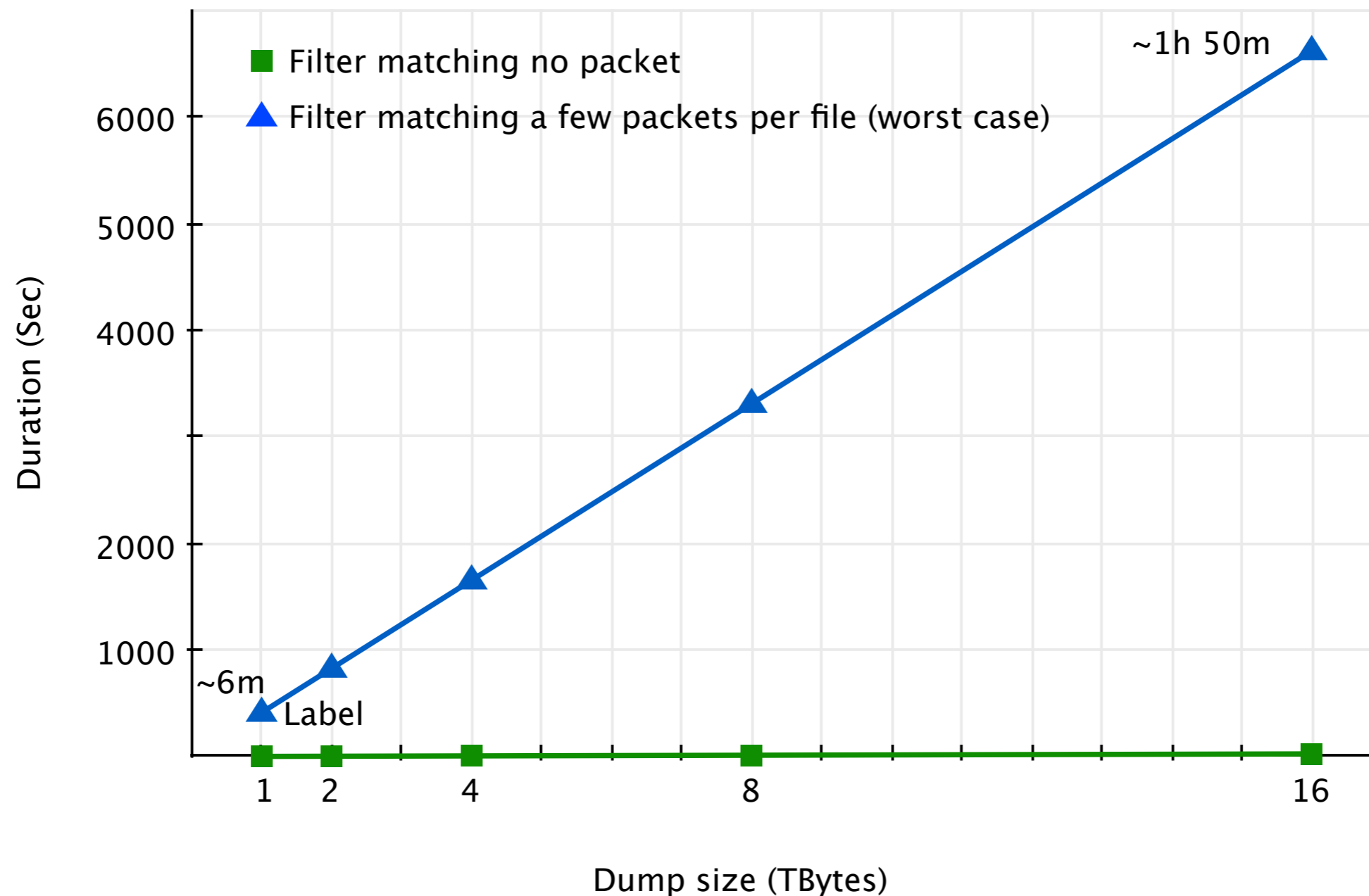
Indexing Scalability

Packet indexing significantly affects performance due to higher per-packet processing costs



Search Scalability

Timeline, blooms and digests dramatically improve packet extraction.

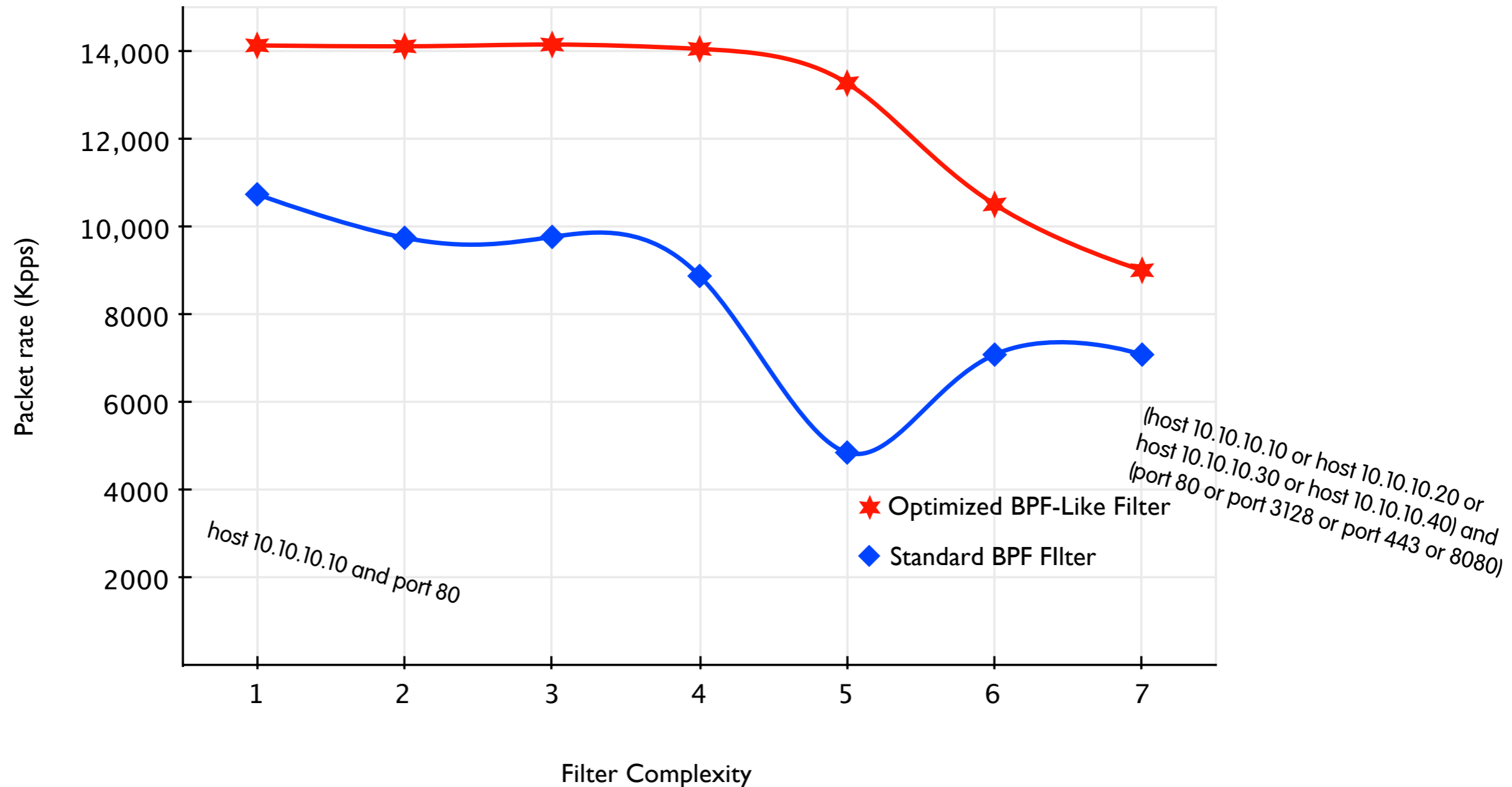


“Fast-BPF”

- BPF is the de-facto format for packet filters.
 - Standard BPF are slow (also during capture).
 - In most cases very simple capture filters are used.
- ➔ faster rules-based filters supporting a subset of the BPF syntax.

BPF vs “Fast-BPF”

Efficiency of packet filtering with sample filters of increasing complexity



Performance Evaluation

- In order to achieve line rate to disk with indexing we need a CPU of at least 2.4 GHz of frequency (minimum per/packet CPU cycles).
- 10 Gbit disk storage configuration:
 - 4 x SSDs or 8 x 10K RPM SATA disks (RAID 1).
 - XFS is the most efficient Linux filesystem.

SW vs HW Timestamps

- It is a popular belief that hardware timestamps are required for accurate monitoring.
- Using the time pulse thread at 10G line rate (67 nsec inter-arrival) we have a different timestamp for each packet (see http://www.ntop.org/pf_ring/who-really-needs-sub-microsecond-packet-timestamps/).
- This accuracy we believe is enough for most people making hw timestamps unnecessary.

Lessons Learnt [1/2]

- Not all PCIe slots are alike: 2 x single-port 10G NIC \neq 1 x dual-port 10G NIC
- Energy efficiency might not be your best friend. Modern CPUs, in particular E5, have strong clock throttle rate optimizations that suddenly change the clock. This might lead to drops unless you set a fix clock speed or use active polling on n2disk to keep frequency high.

Lessons Learnt [2/2]

- Memory bandwidth might become a bottleneck. Use Sandy/Ivy bridge systems or better to maximize the bandwidth.
- CPU clock rate is important (≥ 2.4 GHz). Better to use a cheap E3 (E3-1230 costs ~200\$) than a costly E5 with lower clock.
- NUMA does not help: 8 cores are enough (4 if indexing is not used).
- At 10G n2disk writes ~1.25 GB/sec so disk space is not an option.

Using n2disk in Real Life

- Microburst detection: we have developed a companion application that based on pcap traces identifies microbursts.
- Creation of a “Time Machine” for keeping on disk the recent network communications.
- Using DNA/Libzero in zero-copy run n2disk while generating flows with nProbe.

Future Work

- We have tested multi-10 Gbit packet to disk on a single box. This is feasible, but memory bandwidth can become a bottleneck for transferring packets network->memory->disk.
- We have done some preliminary work on using GPUs for “real” indexing+packet compression: the initial results are promising.

n2disk Availability

- <http://packages.ntop.org>