10 Gbit Line Rate Packet-to-Disk Using n2disk

ntop

Registro.it

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.

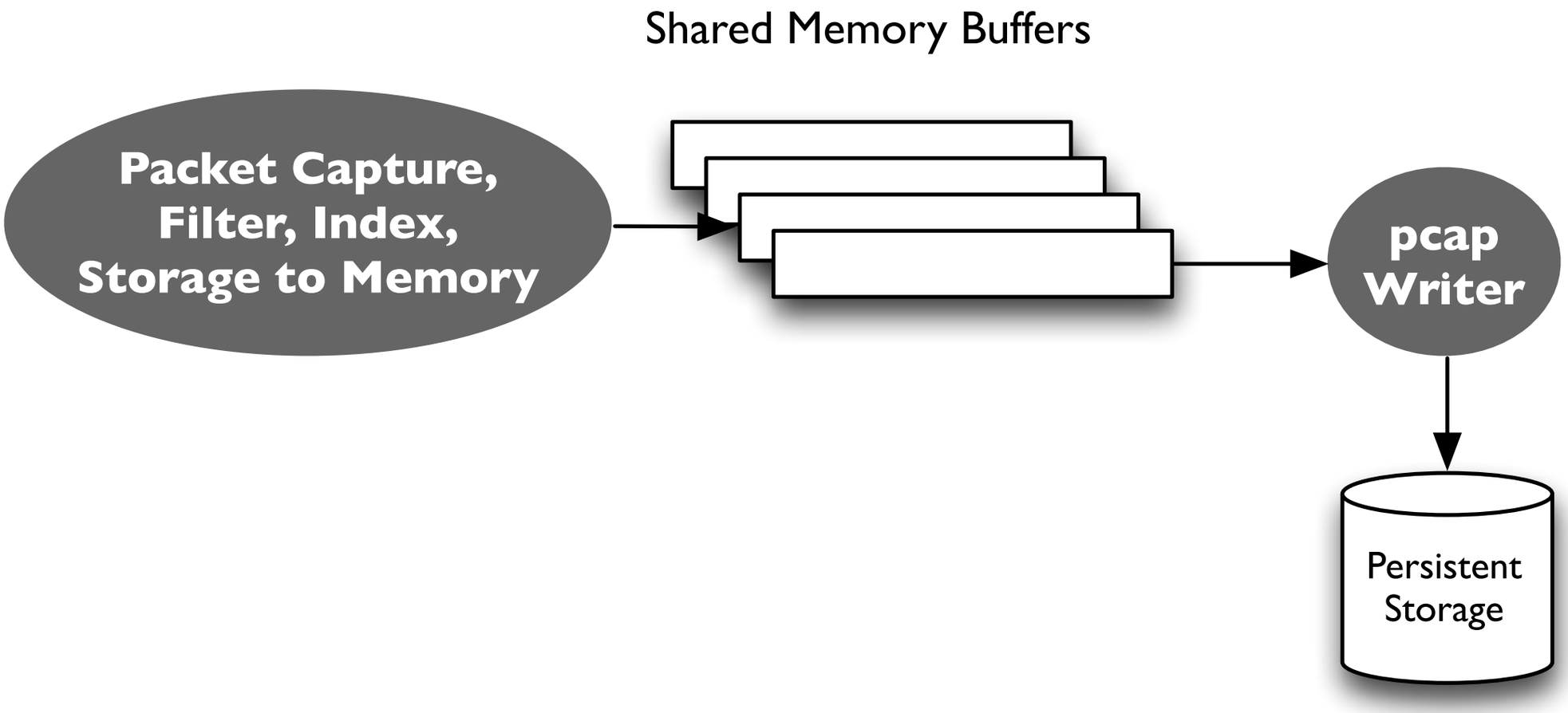
Motivations [2/2]

- 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.

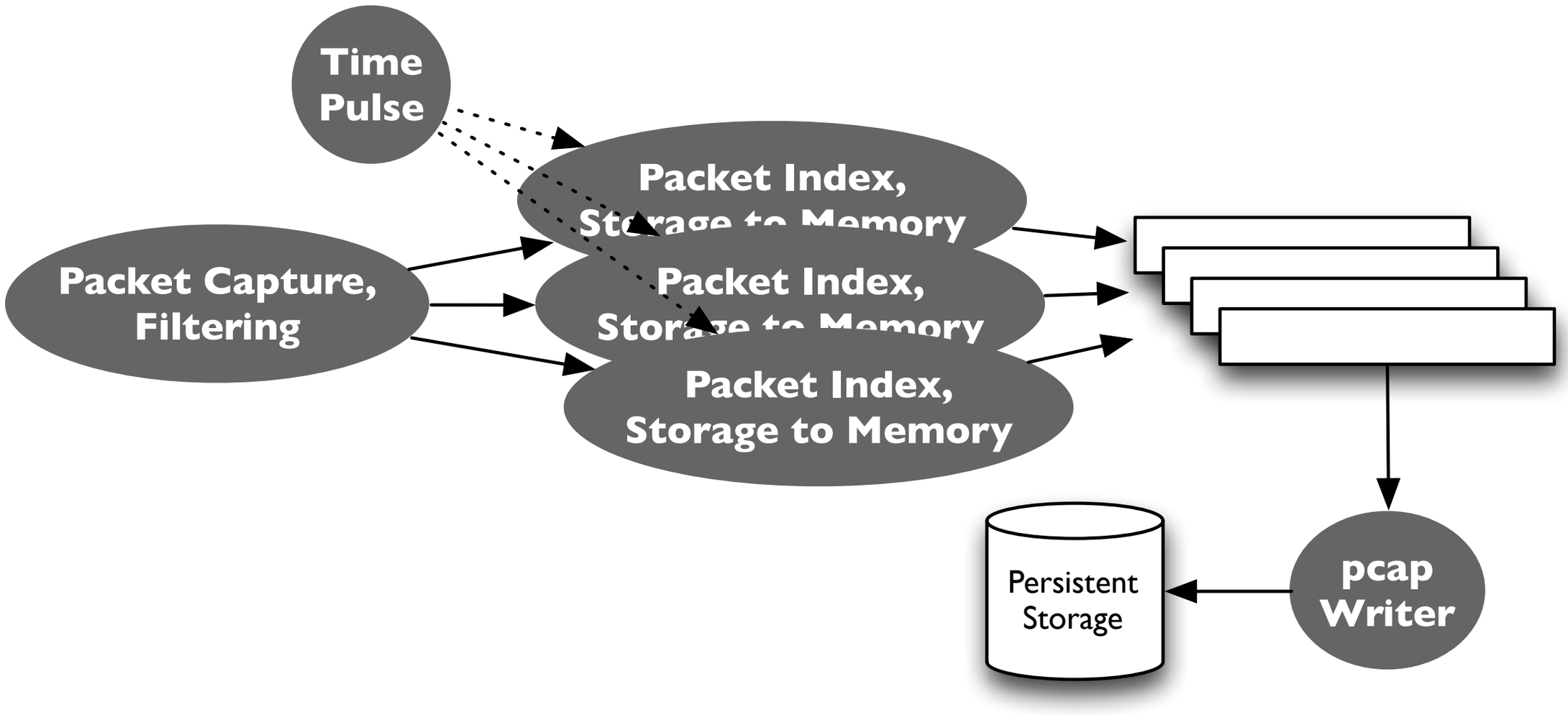
n2disk

- 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

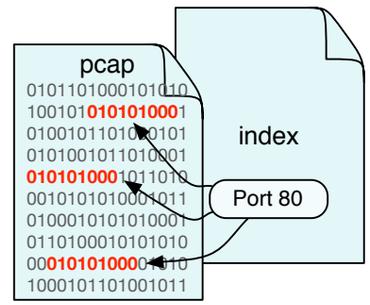


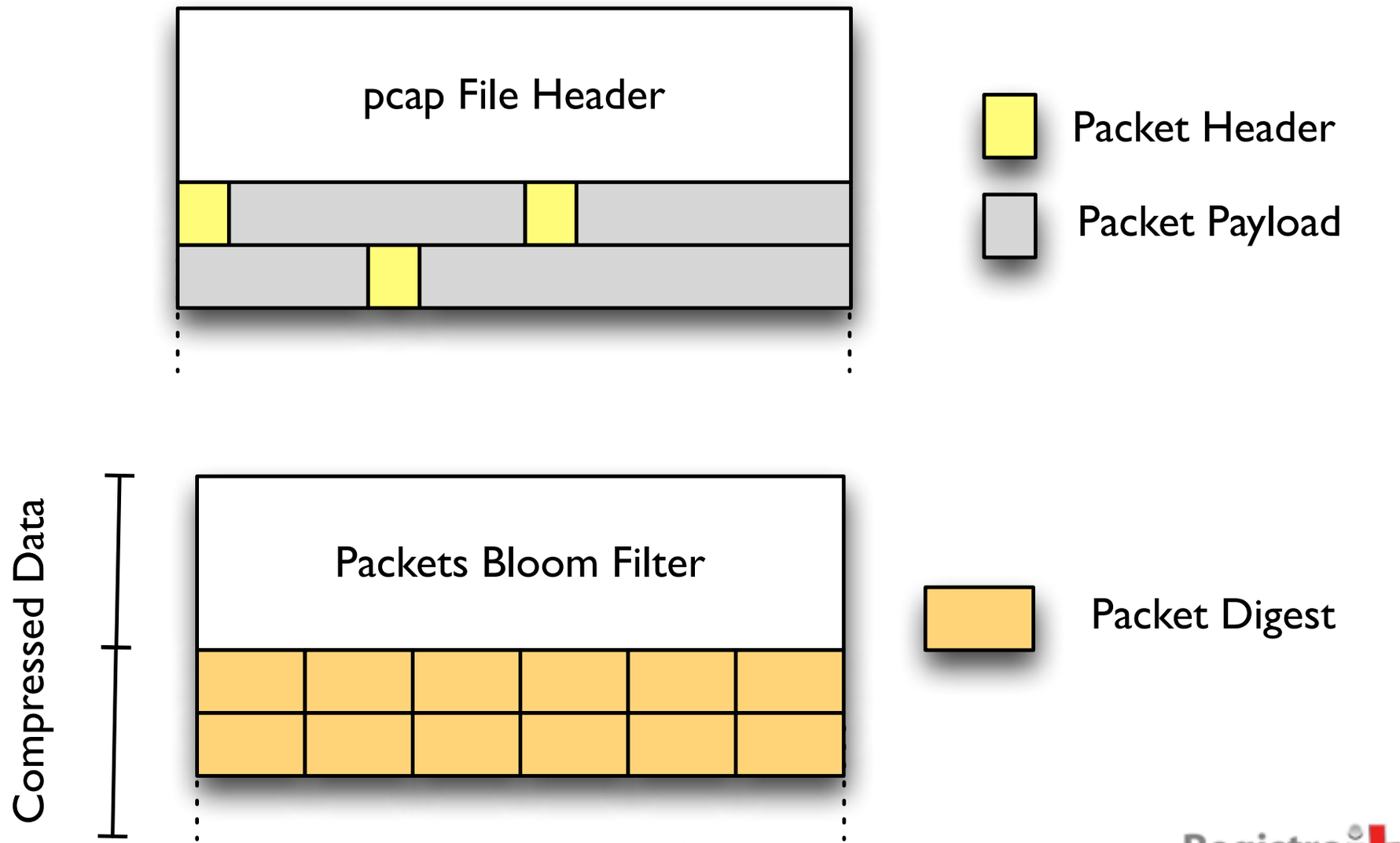
Multi-Threaded Mode



Index

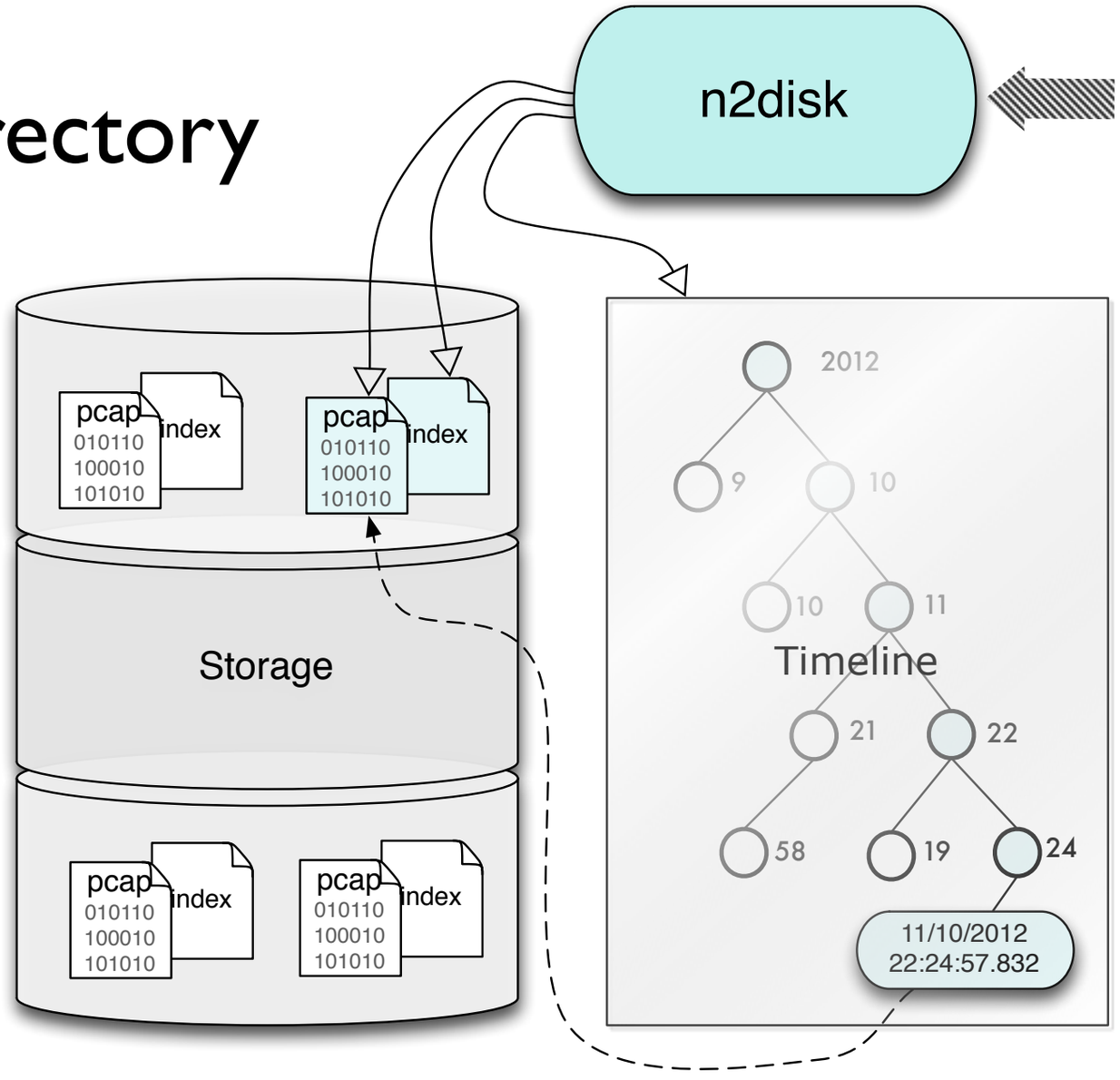
Every pcap file comes with an optional companion index file





Timeline

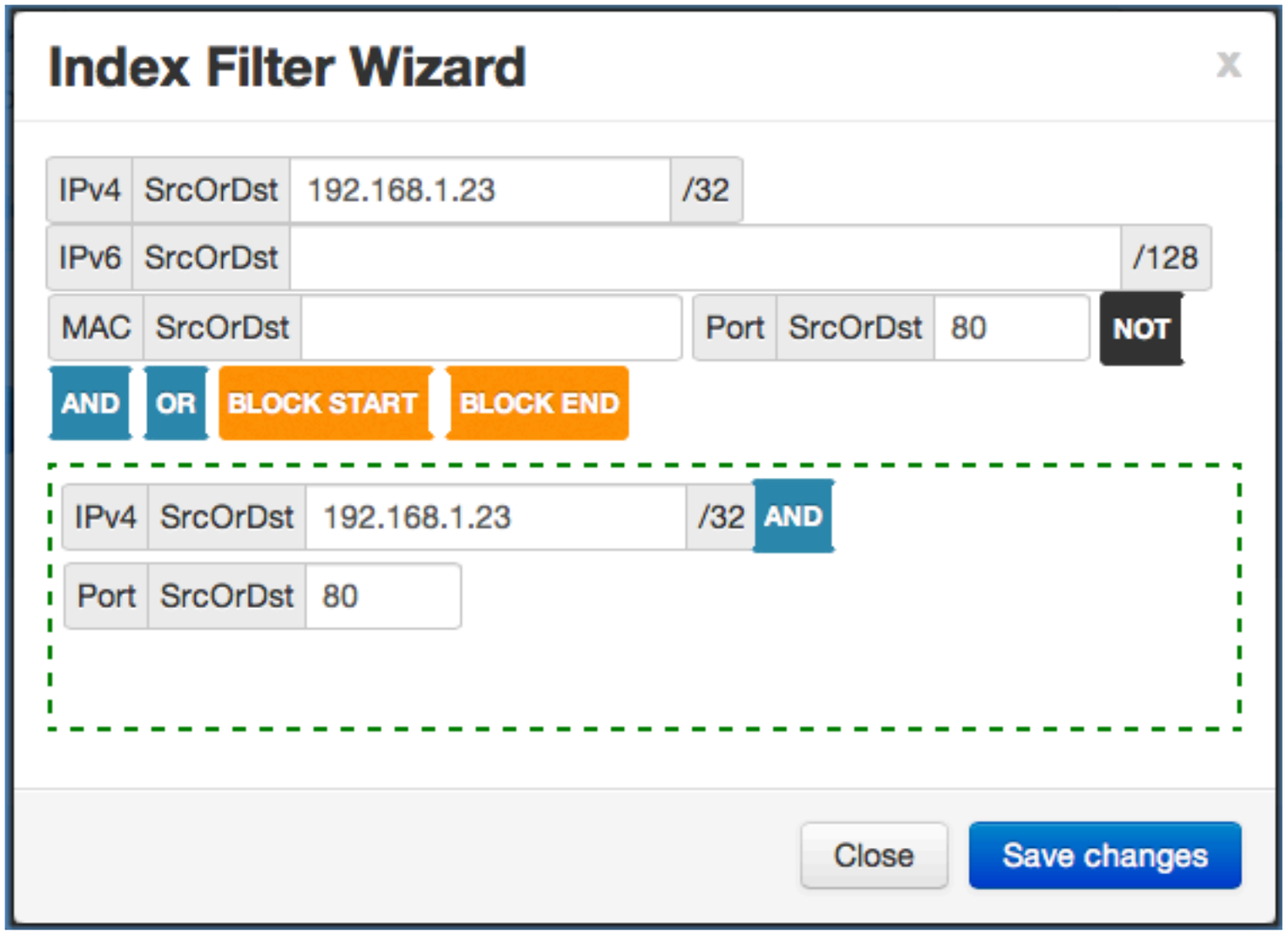
A time-ordered directory tree maintained by n2disk to enable time-based packet extraction



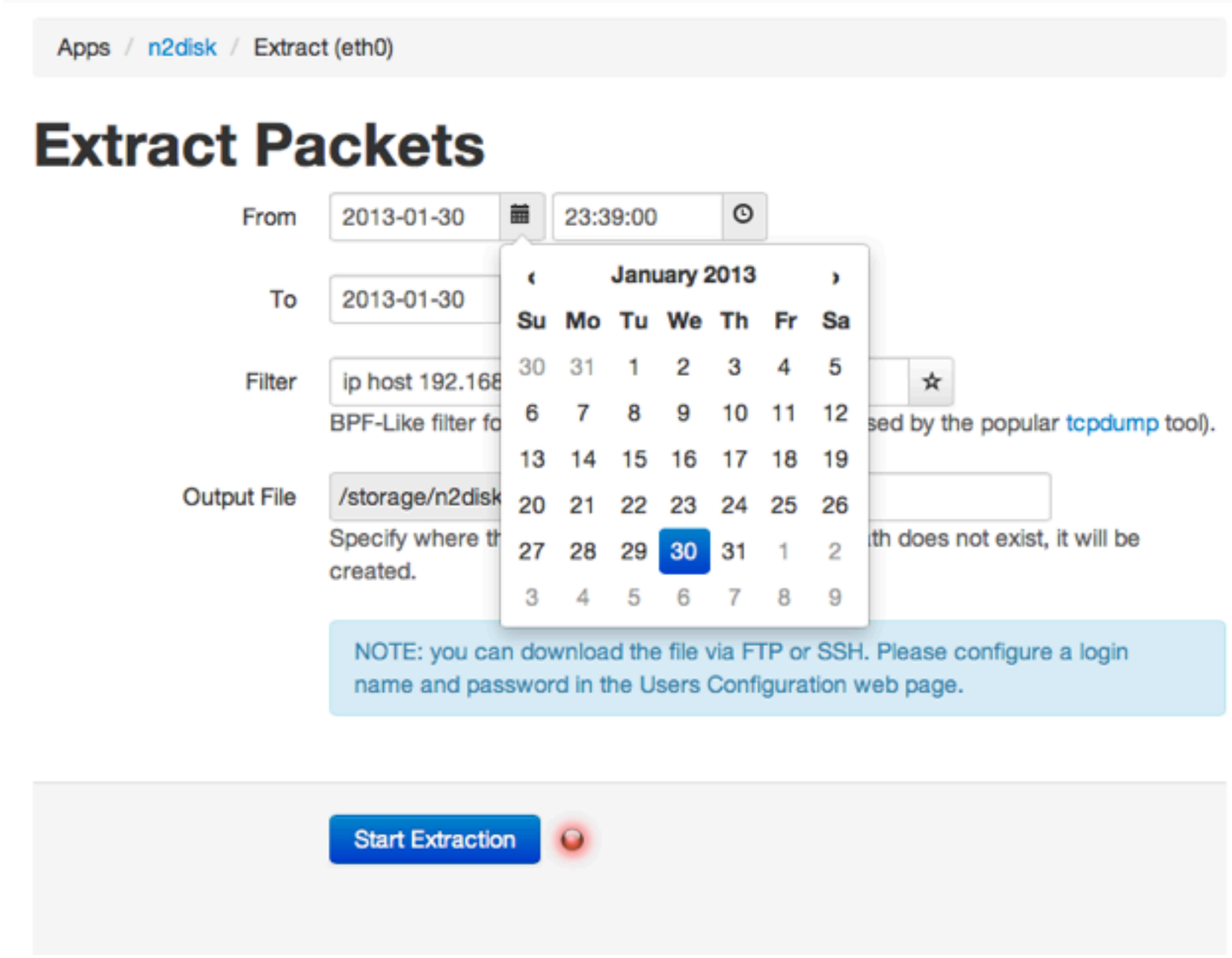
nBox

- 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.

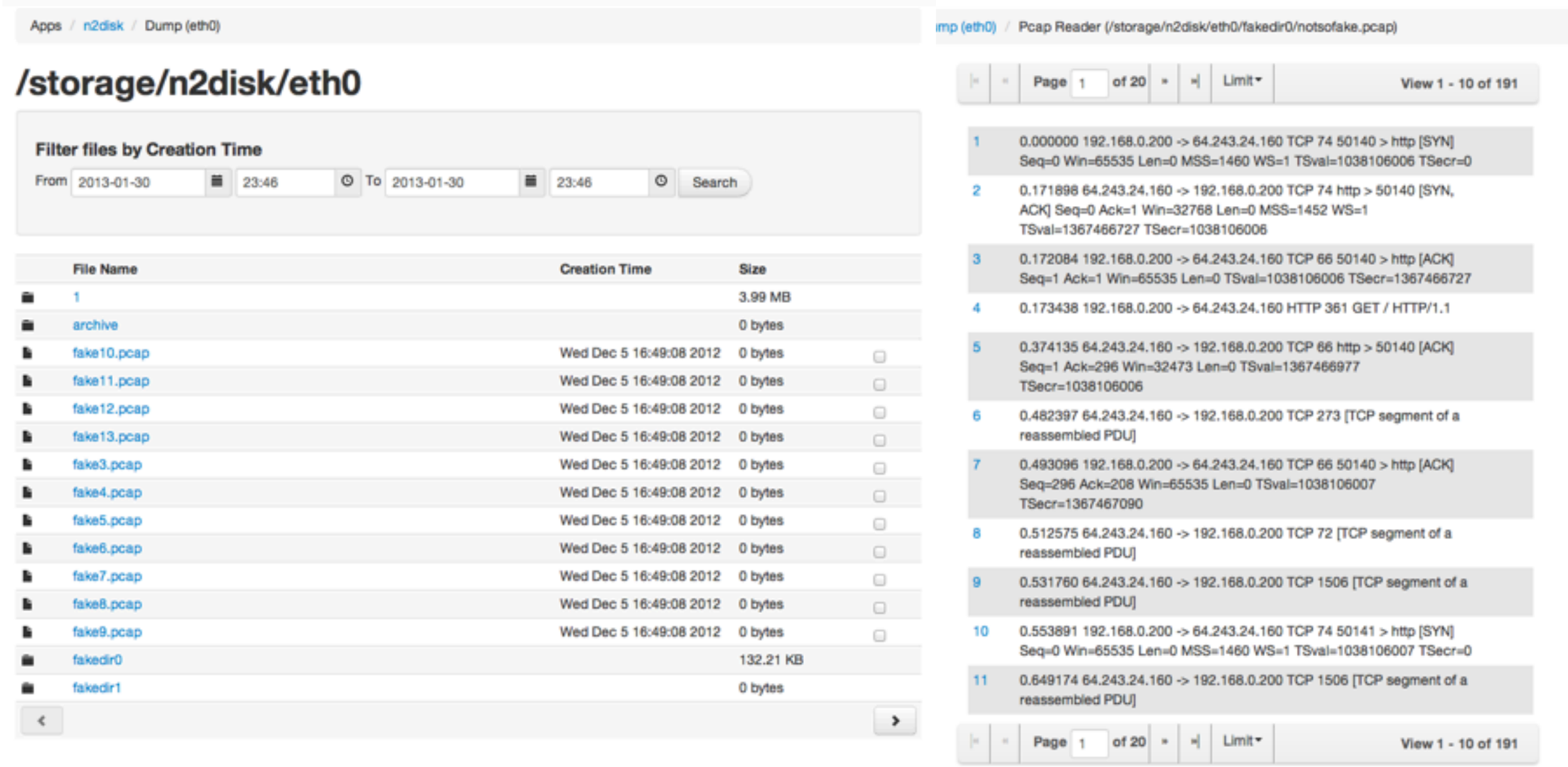
nBox: Filter Wizard



nBox: Packet Extraction

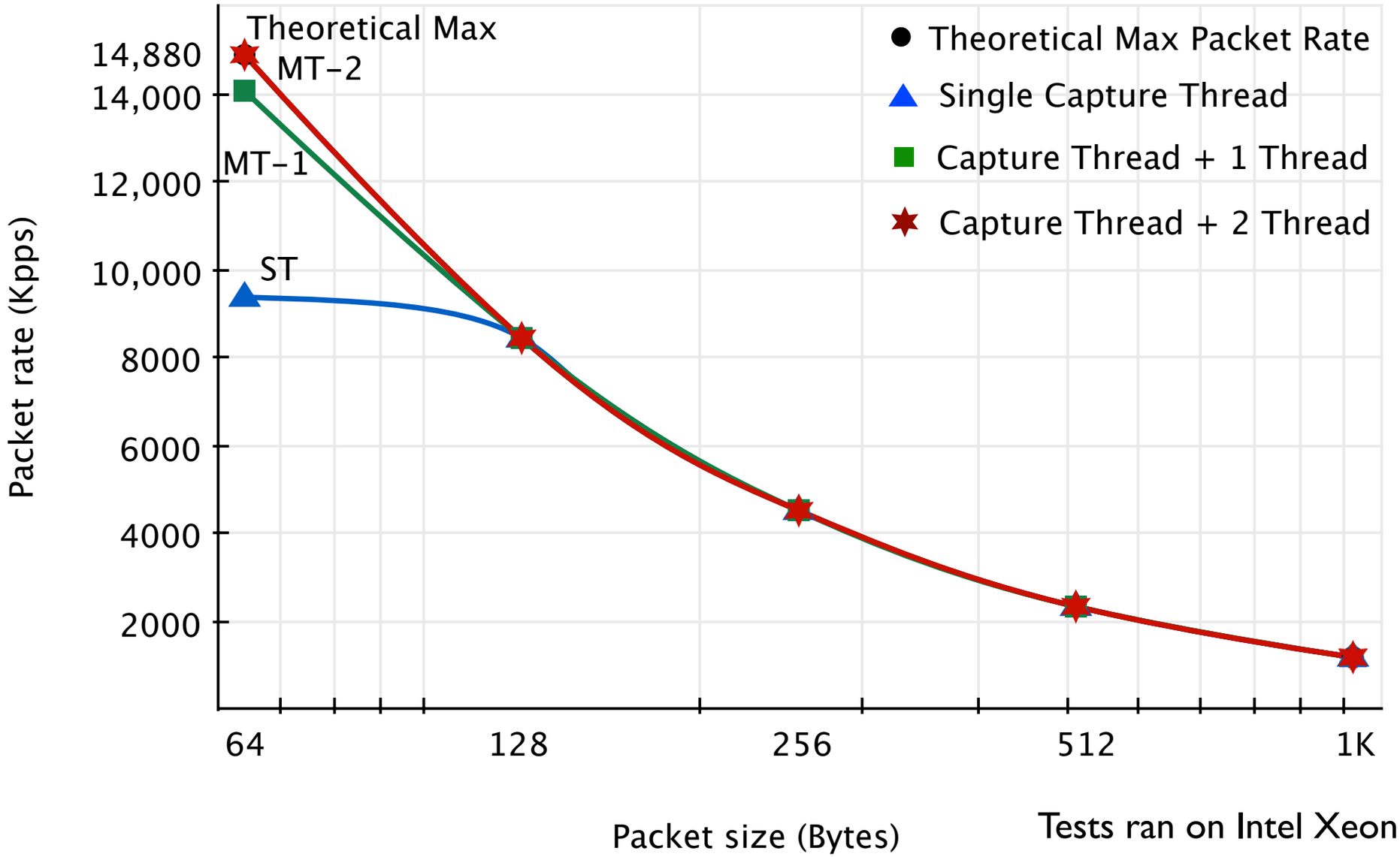


nBox: Packet Browsing



Dump Scalability

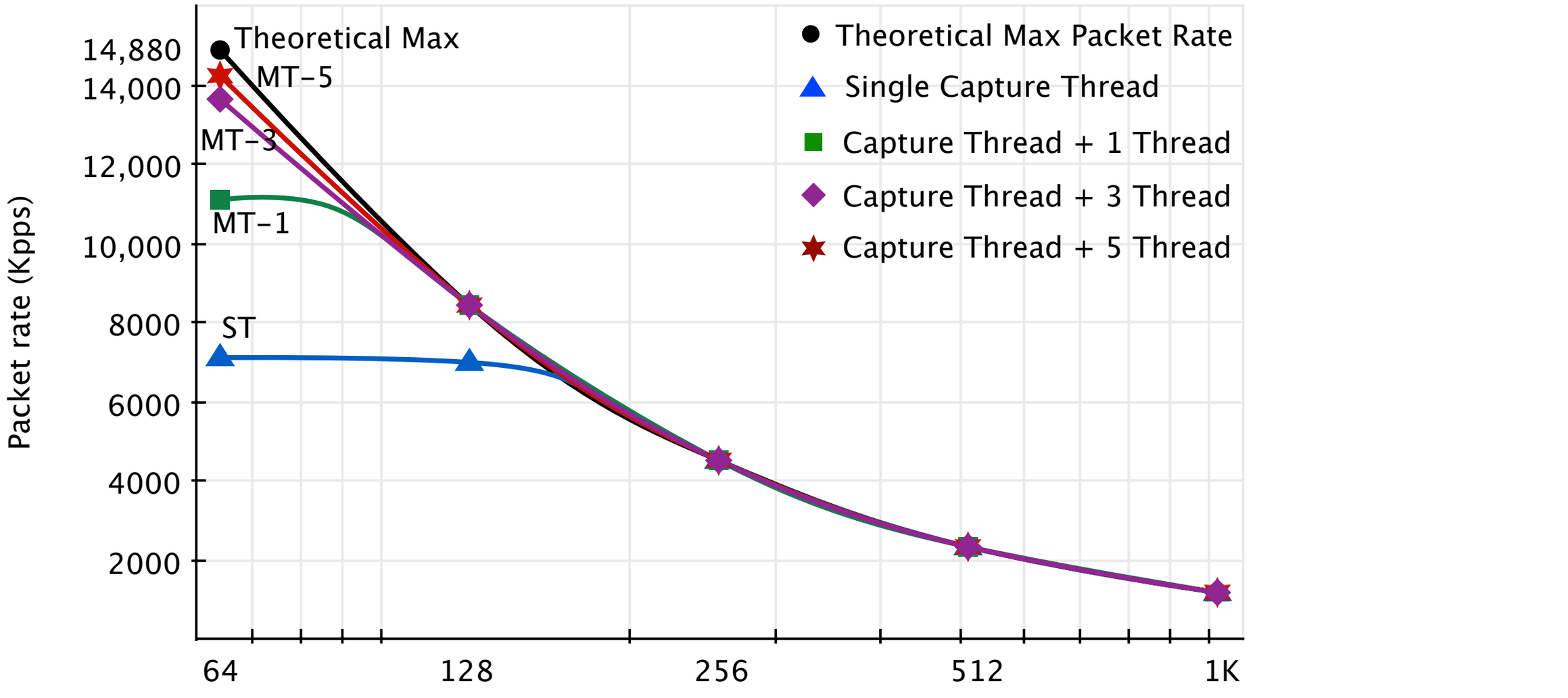
A multithreaded design is a key factor for granting line-rate on low-frequency systems.



Tests ran on Intel Xeon E5-2650 @ 2.00 GHz

Indexing Scalability

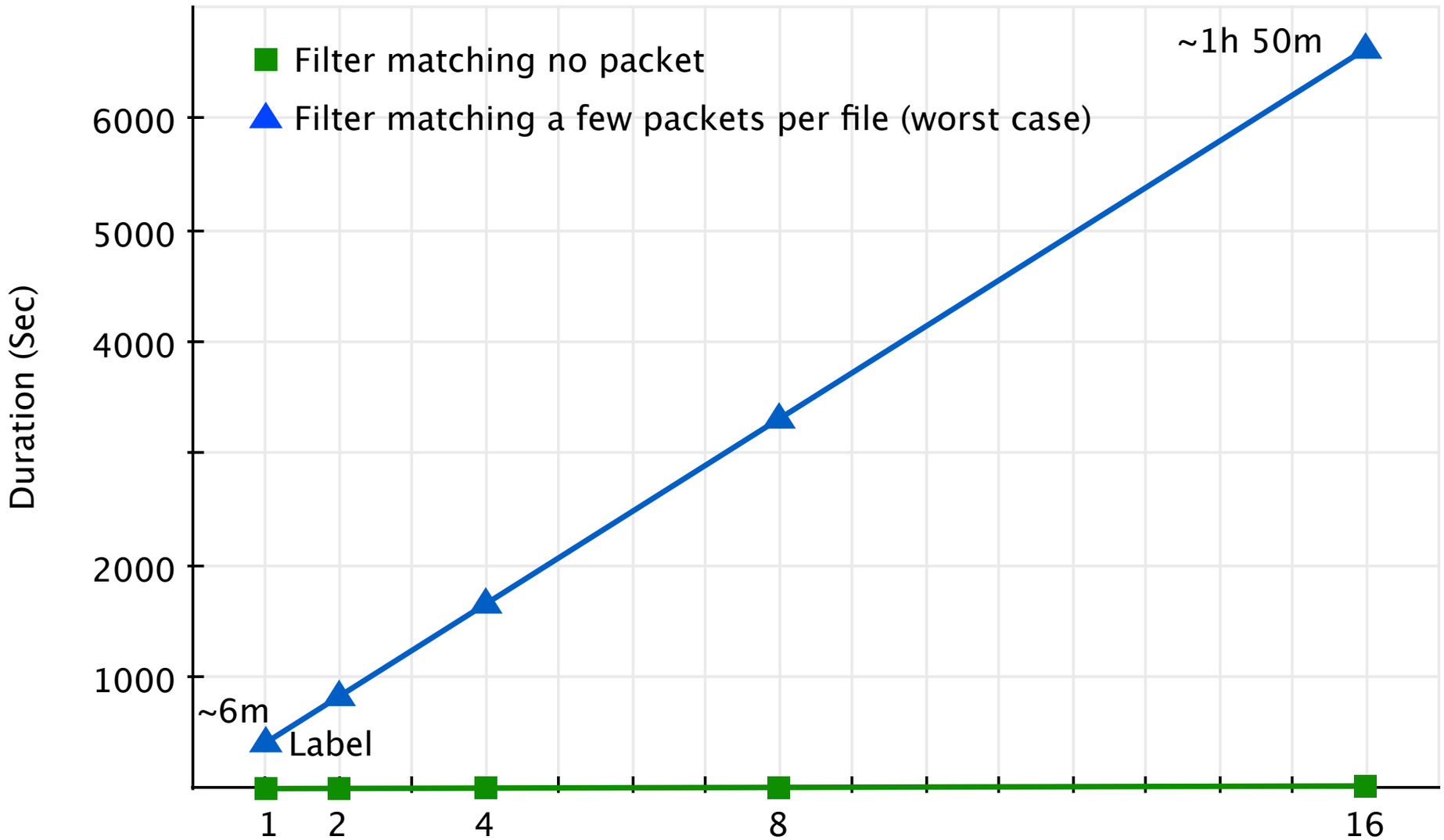
Packet indexing significantly affects performance due to higher per-packet processing costs



Tests ran on Intel Xeon E5-2650 @ 2.00 GHz

Search Scalability

Timeline, blooms and digests dramatically improve packet extraction.



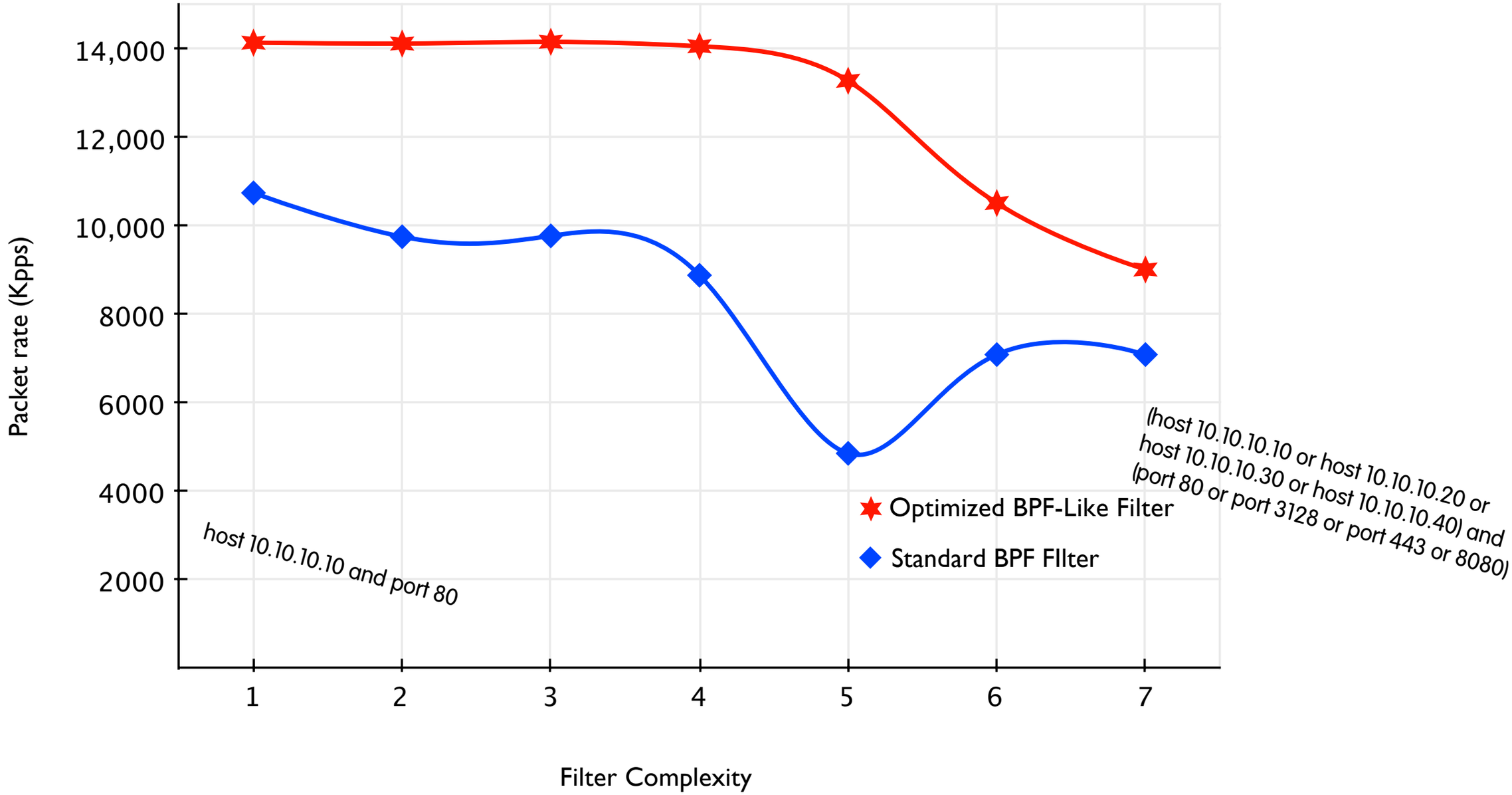
“Fast-BPF”

- 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 require 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 != 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$) that 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