Detecting Very Large Sets Of Referenced Files At 40/100 Gbe, Especially Mp4 Files

By

Adrien Larbanet, Jonas Lerebours and Jean Pierre David

Presented At

The Digital Forensic Research Conference

DFRWS 2015 USA Philadelphia, PA (Aug 9th - 13th)

DFRWS is dedicated to the sharing of knowledge and ideas about digital forensics research. Ever since it organized the first open workshop devoted to digital forensics in 2001, DFRWS continues to bring academics and practitioners together in an informal environment. As a non-profit, volunteer organization, DFRWS sponsors technical working groups, annual conferences and challenges to help drive the direction of research and development.

http:/dfrws.org

Detecting very large sets of referenced files at 40/100 GbE, especially MP4 files

A. Larbanet, J. Lerebours, J.P. David

POLYTECHNIQUE MONTRÉAL

LE GÉNIE EN PREMIÈRE CLASSE

DFRWS

Adrien Larbanet

adrien.larbanet@polymtl.ca

DFRWS 2015, Philadelphia PA, 2015 august 11th

Internet growth

- 2.5 G users and 51.2 EB/mo in 2013.
- Estimated 3.9 G users and 131.6 EB/mo in 2018 (Cisco).
- +50% growth/yr for high-end home connection (Nielsen's law).
- ISP and Internet backbones w/ 10, 40 & 100 Gbps technologies.

Illegal behaviors on the Internet

- Downloading copyright infringing materials.
  - Movies, music, books, video games, etc.
- Illegal intrusions and stealings of confidential documents.
- Sharing child pornography related materials.

What Google, Facebook and Twitter can do

BBC | Sign in | News | Sport | Weather | Shop | Earth | Travel | More | Search

NEWS

Home | Video | World | US & Canada | UK | Business | Tech | Science | Magazine | Entertainment & Arts | Health | More

UK | England | N. Ireland | Scotland | Wales | Politics

UK

Google, Facebook and Twitter to block 'hash lists' of child abuse

10 August 2015 | UK

Web giants Google, Facebook and Twitter have joined forces with a British

Top Stories

Man shot at Ferguson rally charged
Tyrone Harris, who was shot by police in Ferguson during the Michael Brown anniversary events, is charged with assaulting officers, prosecutors say.
1 hour ago

Two killed in Swedish Ikea attack
5 hours ago

Spill turns Colorado river yellow
2 hours ago

Features & Analysis

BBC Website (2015, August 10th). Retrieved from http://www.bbc.com/news/uk-33844124

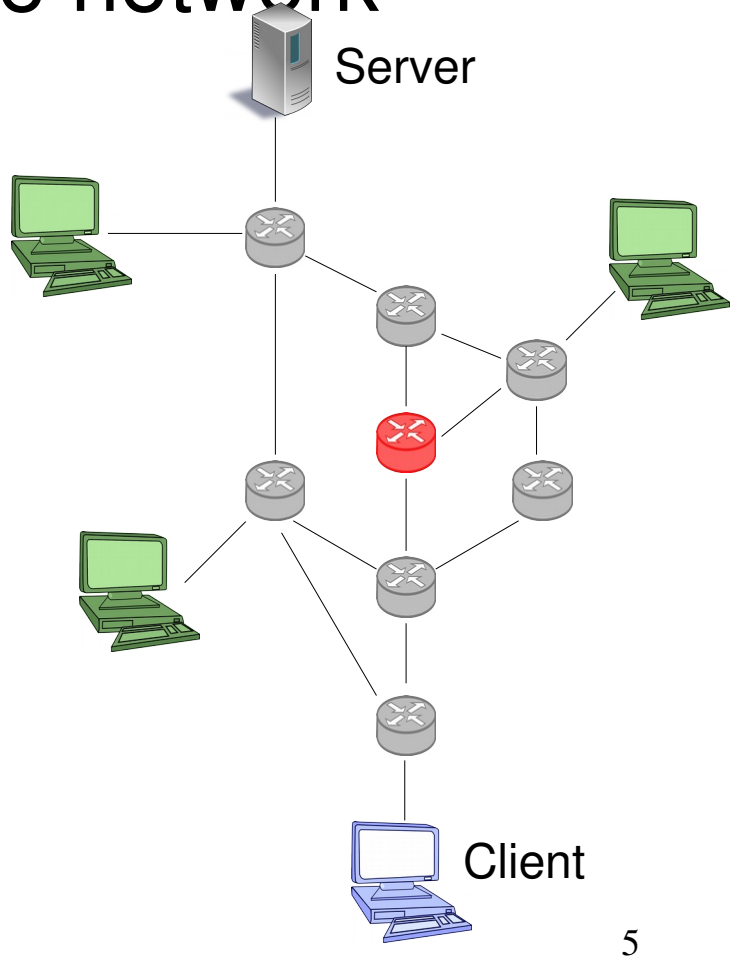
Adrien Larbanet
adrien.larbanet@polymtl.ca

Detection difficulties on the network

- Link speed: up to 10, 40 and 100 Gbps.
- With Ethernet, IP & TCP/UDP protocols :
  - Fragmentation in packets.
  - Packet order is not known.
  - Route is not predetermined.
- Millions of *clean* packets/s (noise).

→Reconstruction is difficult (~impossible).

→Fast detection of known files based on a small subset...

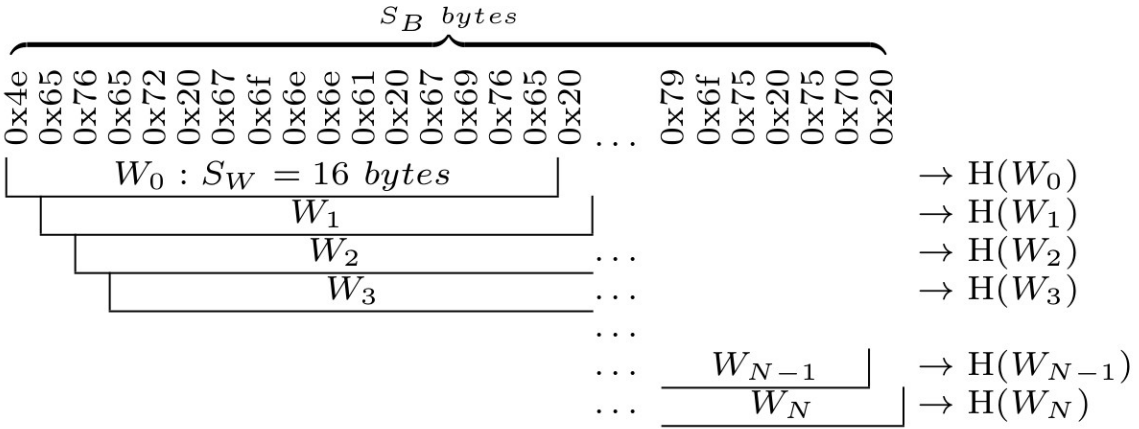


Adrien Larbanet
adrien.larbanet@polymtl.ca

A solution: the max-hashing algorithm

- Presented by David in 2013 [1].
- Focuses on small windows (16B) w/ a local property:

  → Local maximum hash.



$$Final fingerprint : F = \max_{0 \leq n \leq N} (\mathrm{H}(W_n)) \quad with \quad N = S_B - S_W$$

Adrien Larbanet
adrien.larbanet@polymtl.ca

Max-hashing – Principle

Basic referencing

- Fragment the file in blocks.
- Compute a hash for each 16B window on the blocks.
- Store every local maximum into database.

Detection

- Capture network packets
- Compute a hash for each 16B window on the payloads.
- Compare the local maxima w/ the referenced fingerprints.

→ Packets are independent: task is parallelizable.

Adrien Larbanet
adrien.larbanet@polymtl.ca

Basic referencing with Max-Hashing

Once upon a midnight dreary, while I pondered weak and weary, Over many a quaint and curious volume of forgotten lore,While I nodded, nearly napping, suddenly there came a tapping,As of some one gently rapping, rapping at my chamber door.`''Tis some visitor,' I muttered, `tapping at my chamber door -Only this, and nothing more.' Ah, distinctly I remember it was in the bleak December,And each separate dying ember wrought its ghost upon the floor.Eagerly I wished the morrow; -vainly I had sought to borrowFrom my books surcease of sorrow - sorrow for the lost Lenore -For the rare and radiant maiden whom the angels name Lenore -Nameless here for evermore. And the silken sad uncertain rustling of each purple curtainThrilled me - filled me with fantastic terrors never felt before;So that now, to still the beating of my heart, I stood repeating`''Tis some visitor entreating entrance at my chamber door -Some late visitor entreating entrance at my chamber door; -This it is, and nothing more,'

Edgar A. Poe, The Raven

Adrien Larbanet
adrien.larbanet@polymtl.ca

Basic referencing with Max-Hashing

- Fragment document into blocks.

- Once upon a midnight dreary, while I pondered weak and weary, Over many a quaint and curious volume of forgotten lore,While I nodded, nearly napping, suddenly there came a tapping,As of some one gently rapping, rapping at my chamber door.`'Tis some visitor,' I muttered, `tapping at my chamber door -Only this, and nothing more.'

- Ah, distinctly I remember it was in the bleak December,And each separate dying ember wrought its ghost upon the floor.Eagerly I wished the morrow; - vainly I had sought to borrowFrom my books surcease of sorrow - sorrow for the lost Lenore -For the rare and radiant maiden whom the angels name Lenore -Nameless here for evermore.

- And the silken sad uncertain rustling of each purple curtainThrilled me - filled me with fantastic terrors never felt before;So that now, to still the beating of my heart, I stood repeating`'Tis some visitor entreating entrance at my chamber door -Some late visitor entreating entrance at my chamber door; -This it is, and nothing more,'

Edgar A. Poe, The Raven

Adrien Larbanet
adrien.larbanet@polymtl.ca

Basic referencing with Max-Hashing

- Fragment document into blocks.
- Hash every 16-character wide windows.
- Keep local maximum in every block.

  → Are all 16-character strings equi-probable in language ?

- Some are *common.*
- Some are *unique.*

- Once upon a midnight dreary, while I pondered weak and weary, Over many a quaint and curious volume of forgotten **lore,While I nod**ded, nearly napping, suddenly there came a tapping,As of some one gently rapping, rapping at my chamber door.`'Tis some visitor,' I muttered, `tapping at my chamber door -Only this, and nothing more.'

- Ah, distinctly **I remember it wa**s in the bleak December,And each separate dying ember wrought its ghost upon the floor.Eagerly I wished the morrow; - vainly I had sought to borrowFrom my books surcease of sorrow - sorrow for the lost Lenore -For the rare and radiant maiden whom the angels name Lenore -Nameless here for evermore.

- And the silken sad uncertain rustling of each purple curtainThrilled me - filled me with fantastic terrors never felt before;So that now, to still the beating of my heart, I stood repeating`'Tis some visitor entreating entrance at my chamber door -Some late visitor entreating entrance at my cha**mber door; -This** it is, and nothing more,'

Edgar A. Poe, The Raven

Adrien Larbanet
adrien.larbanet@polymtl.ca

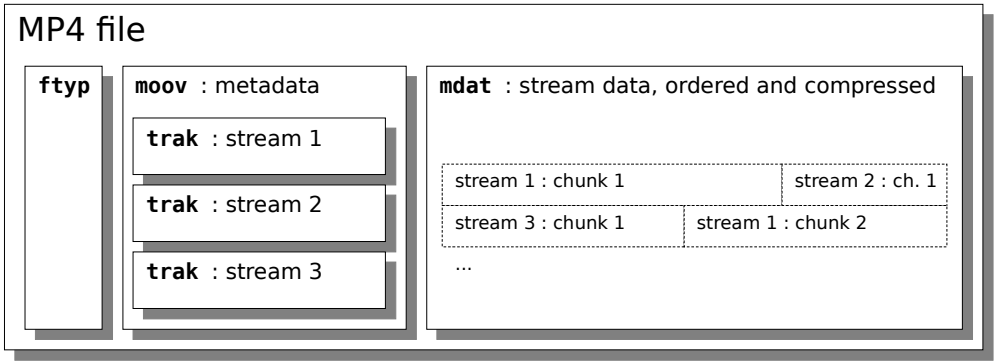
Referencing files – Problems with MP4+H.264

- Using CCV video DB [2]: 8073 MP4 files.

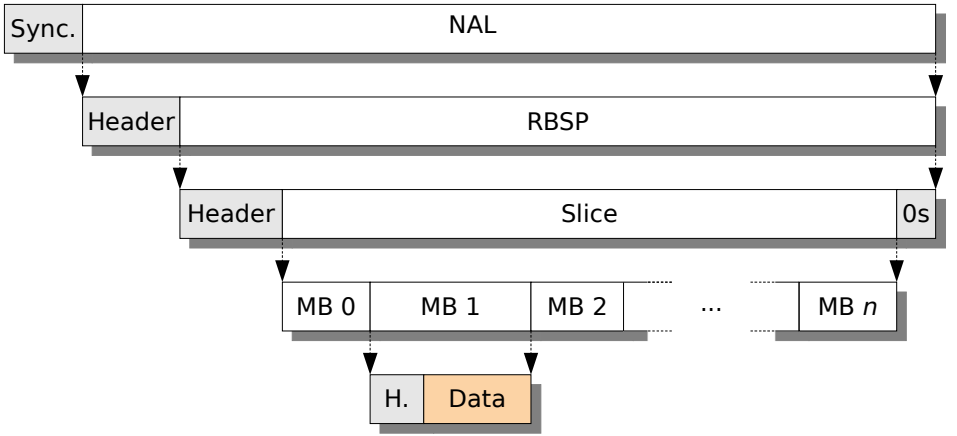
1. Keep 512 local maxima per file:
   - Over 4 Millions fingerprints.
   - 371 redundancies.
   - 243 files implied.
2. With detection setup:
   - Over 500 Millions comparisons with fingerprints in DB.
   - 18000 false-positives.

Referencing files – MP4/H.264 file structure

MP4



H.264



Adrien Larbanet
adrien.larbanet@polymtl.ca

Advanced referencing for MP4/H.264 files

- Modified free and open-source software Ffmpeg:
  - Get the position of the “high-entropy” data field for each Macro-Blocks.
  - When referencing, only keep local maximum computed on these segments.

1. Refencing 8073 files:
   - No redundancies.
2. Detection of half the DB:
   - No false-positive.

Adrien Larbanet
adrien.larbanet@polymtl.ca

Let's detect !

Detection – GPU architecture

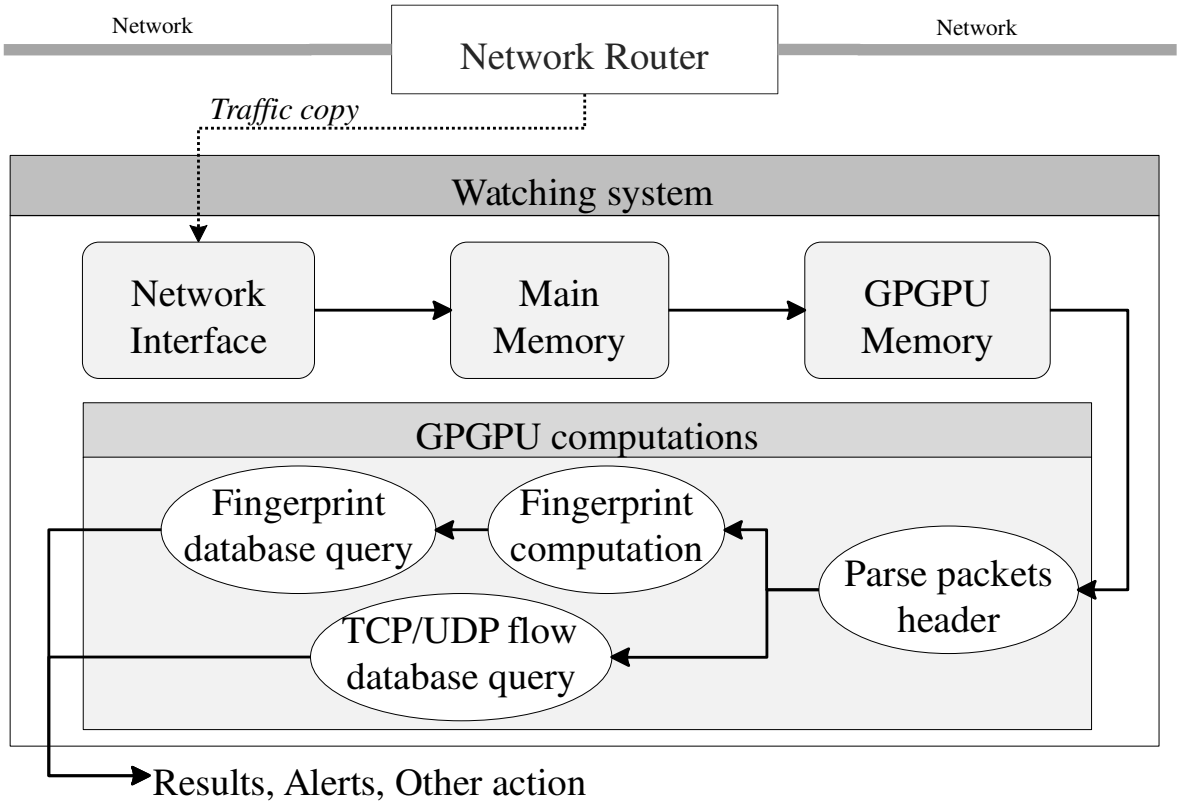
- *SIMD* based.
  - Parallelizable task on independent data.
- Greater computation power (than CPU):

  > 4Tflops (Single-precision)
  for high-end GPU.
- Cheaper and easier to program than FPGA.

NVIDIA GK110 Multiprocessor

Adrien Larbanet
adrien.larbanet@polymtl.ca

Max-Hashing with GPU, system overview

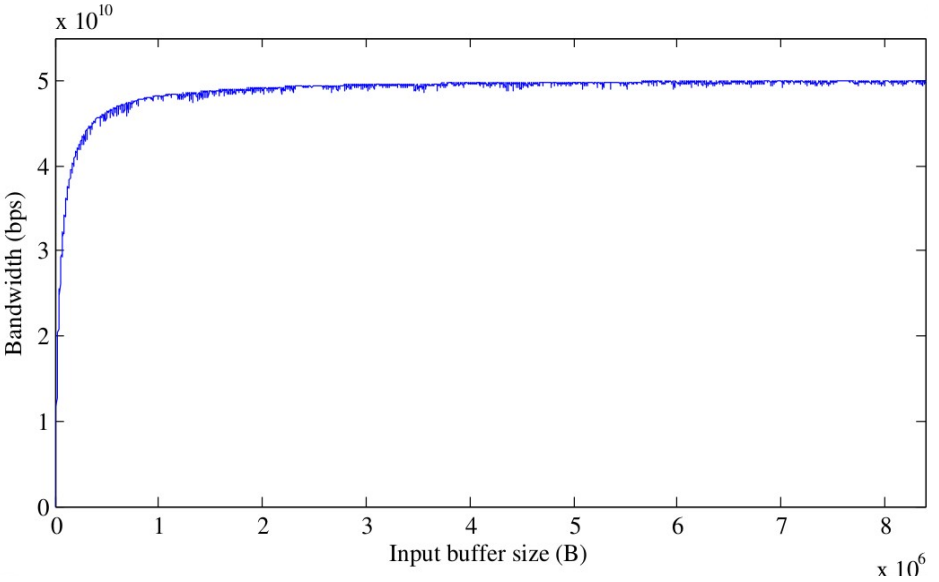


- Goals:
  - Detect Ethernet packets carrying known content.
  - Flag all the packets belonging to a suspicious stream.

Adrien Larbanet
adrien.larbanet@polymtl.ca

Transfer Ethernet packets over PCIe 2.0

- From DDR3 main memory.
- To onboard GDDR5.
- Through PCIe 2.0 (16x).
  - 64 Gbps in theory.
- Using CUDA DMA.

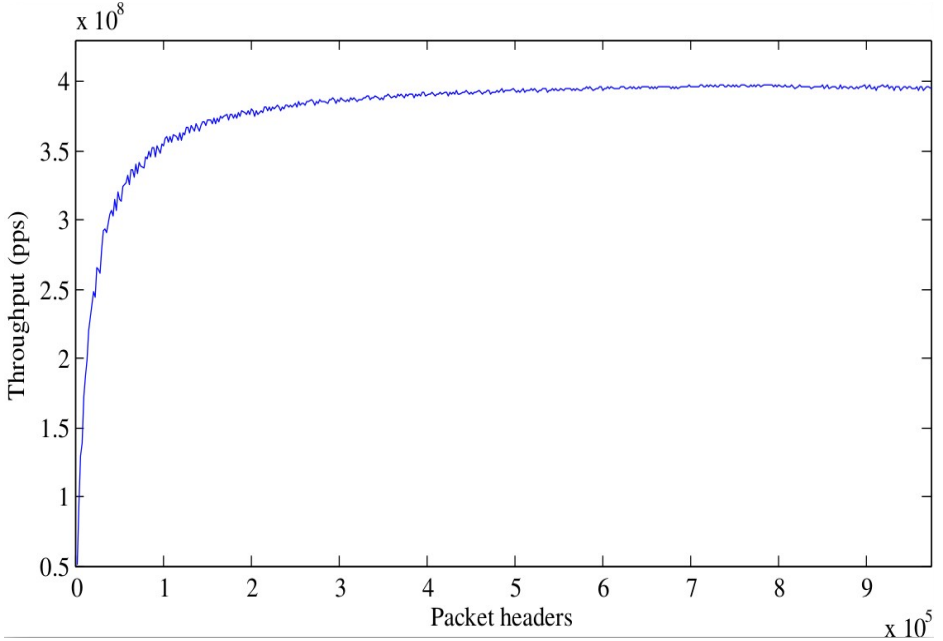


→ Up to 50 Gbps.

Adrien Larbanet
adrien.larbanet@polymtl.ca

Parsing headers on GPU

- *Kernel* tasks for each packet:
  - Getting payload offset and length (store these in memory).
  - Extracting TCP/UDP-IP information for latter analysis.



400 Millions packets per second

Adrien Larbanet
adrien.larbanet@polymtl.ca

Max-Hashing on GPU

- *Kernel* tasks:
  - Get payload position and length.
  - Hash payload and extract 4 local maximum hashs.
- Benchmark setup :
  - 768 MB of random data hashed by 524,288 threads.

→119.9 Gbps, producing 40 millions fingerprints per second.

Adrien Larbanet
adrien.larbanet@polymtl.ca

Fingerprints and TCP/UDP flows databases

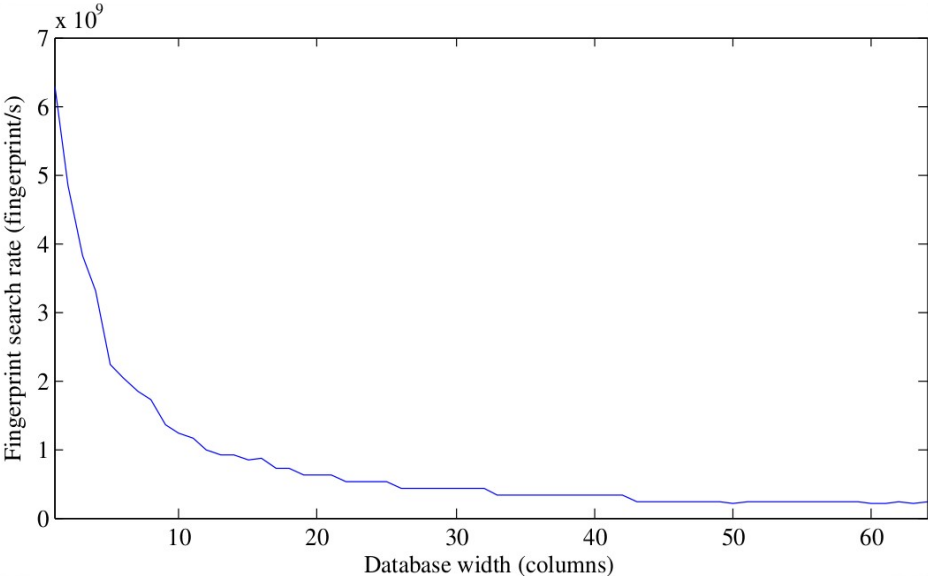
- Hashtable :
  - 2^p rows and *c* columns.
  - *p* LSb of fingerprint as index in the hashtable.
  - For each fingerprint to compare, *c* threads are launched.
- Example with 8-bit fingerprints (p=4):
  - Is 0x22 in the DB ?

	0	*1*	*2*	...	*c*
0	0xA0	0x30	0	...	0
1	0x21	0	0	...	0
2	0xE2	0x22	0xF2	...	0x32
3	0	0	0	...	0
4	0x54	0x24	0xB4	...	0
5	0xC5	0	0	...	0
...	...	...	...	...	...
15	0xEF	0x0F	0	...	0

Adrien Larbanet
adrien.larbanet@polymtl.ca

Fingerprint Database embedded in GPU mem.

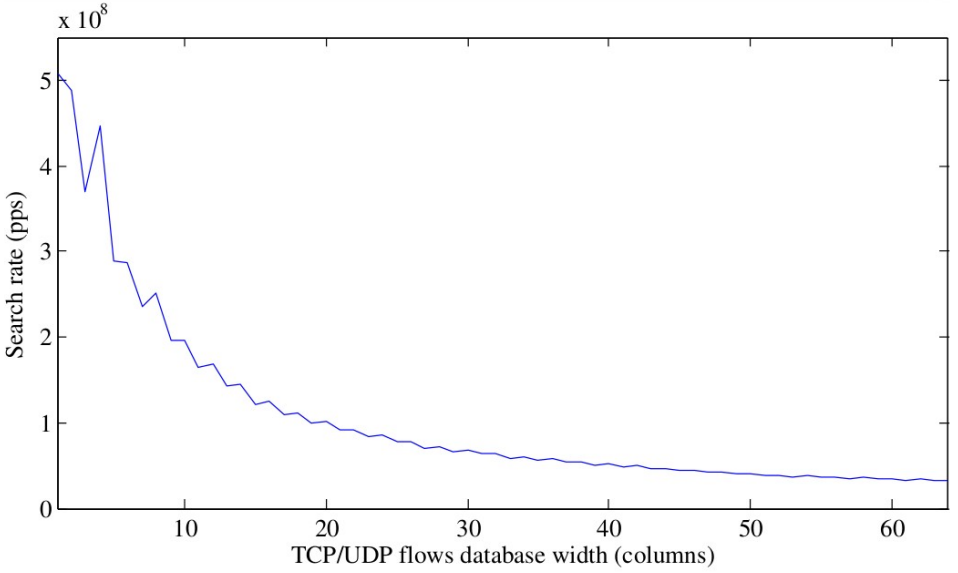
- Hashtable configuration:
  - Contains referenced fingerprints (64 bits each).
  - 2^{21} rows, c_1 columns.
- Benchmark setup
  - Over 2 millions random fingerprints to search in the DB.
  - c_1 is variable.



Adrien Larbanet
adrien.larbanet@polymtl.ca

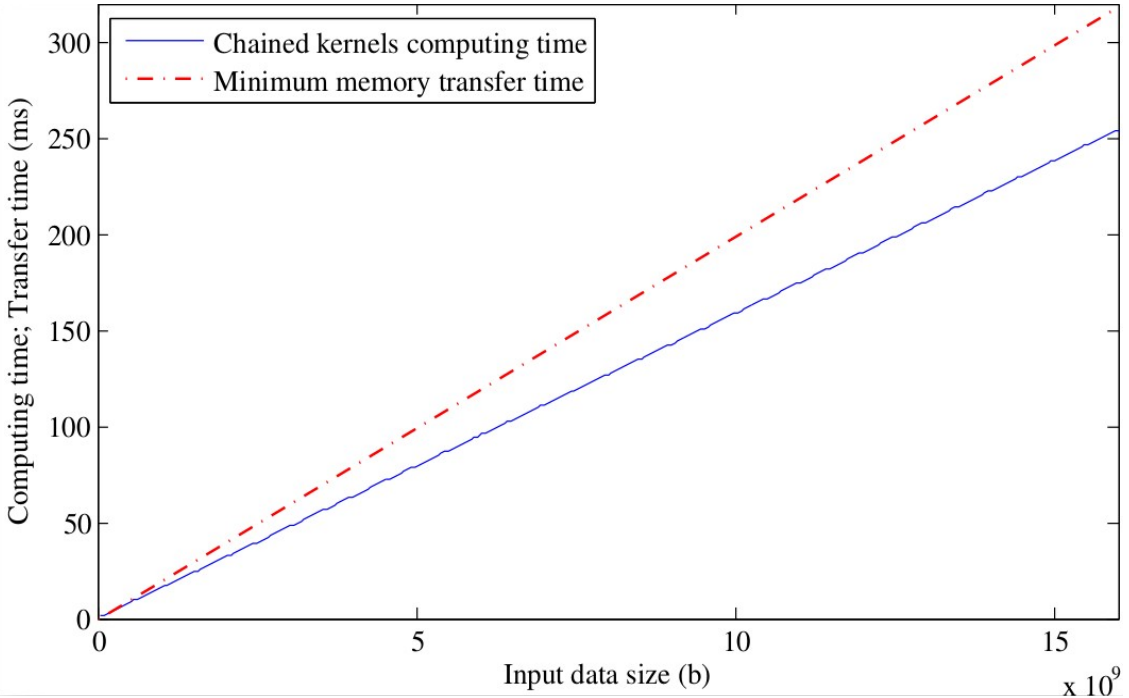
TCP/UDP-IP flows database in GPU mem.

- Hashtable configuration:
  - Contains IP@ and TCP/UDP port boundaries.
  - 2^{16} rows, c_2 columns.
  - 16-bit index computed from IP@.
- Benchmark setup
  - More than half a million random TCP/UDP-IP flow informations.
  - c_2 is variable.



Adrien Larbanet
adrien.larbanet@polymtl.ca

Chained *kernels*



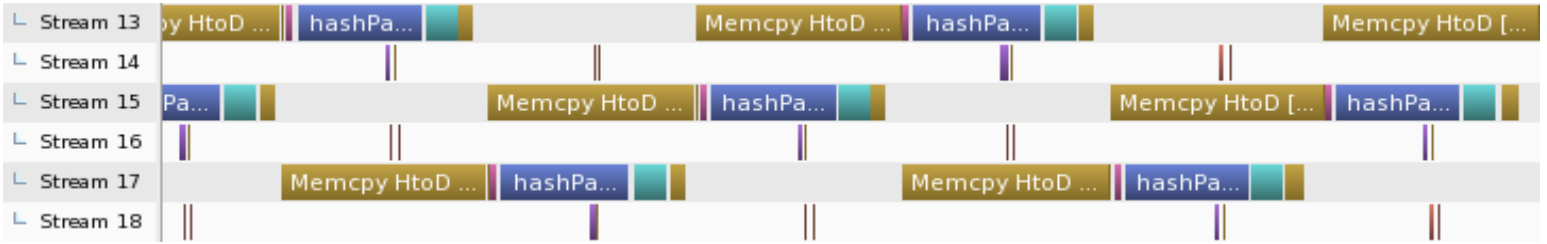
Transfers through PCIe 2.0: **50 Gbps**.
Chained *kernels:* **63 Gbps**.

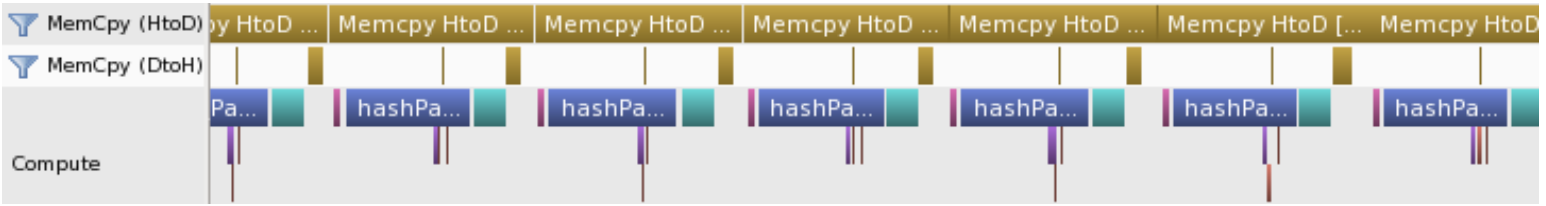
Sequentially: ~25 Gbps...

Pipelining computation and transfer ?

Adrien Larbanet
adrien.larbanet@polymtl.ca

Pipelining – Implementation





Adrien Larbanet
adrien.larbanet@polymtl.ca

Conclusion

Referencing MP4 files

- Basic approach not suitable for heavily formatted files.
- Quick format study reveals position of high-entropy segments.
  - Good candidates for hashing!

Detection on GPU

- Up to 50 Gbps.
- Suitable for 40 GbE.
- Limited by PCIe 2.0...

Adrien Larbanet
adrien.larbanet@polymtl.ca

Discussion - Future work

- Switch to a PCI-e 3.0 x16 compatible GPU:
  - 128 Gbps in theory.
- Multiple GPU setups.
  - Mirror setup.
  - Database dedicated GPU.
  - ...
- Extend dedicated referencing to other video and audio formats:
  - H.265, VP9, etc.

References and acknowledgments

- [1] David, J. P., 2013. Max-hashing fragments for large data sets detection. In: Reconfigurable Computing and FPGAs (ReConFig), 2013 International Conference on. pp. 1–6.
- [2] Jiang, Y.-G., Ye, G., Chang, S.-F., Ellis, D., Loui, A. C., 2011. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In: Proceedings of ACM International Conference on Multimedia Retrieval (ICMR), oral session.

The authors are grateful to NetClean and the Government of Quebec.

Adrien Larbanet
adrien.larbanet@polymtl.ca

Thank you !

Any question ?

Adrien Larbanet
adrien.larbanet@polymtl.ca