Michalis Kallitsis, Merit Network, Inc. / University of Michigan

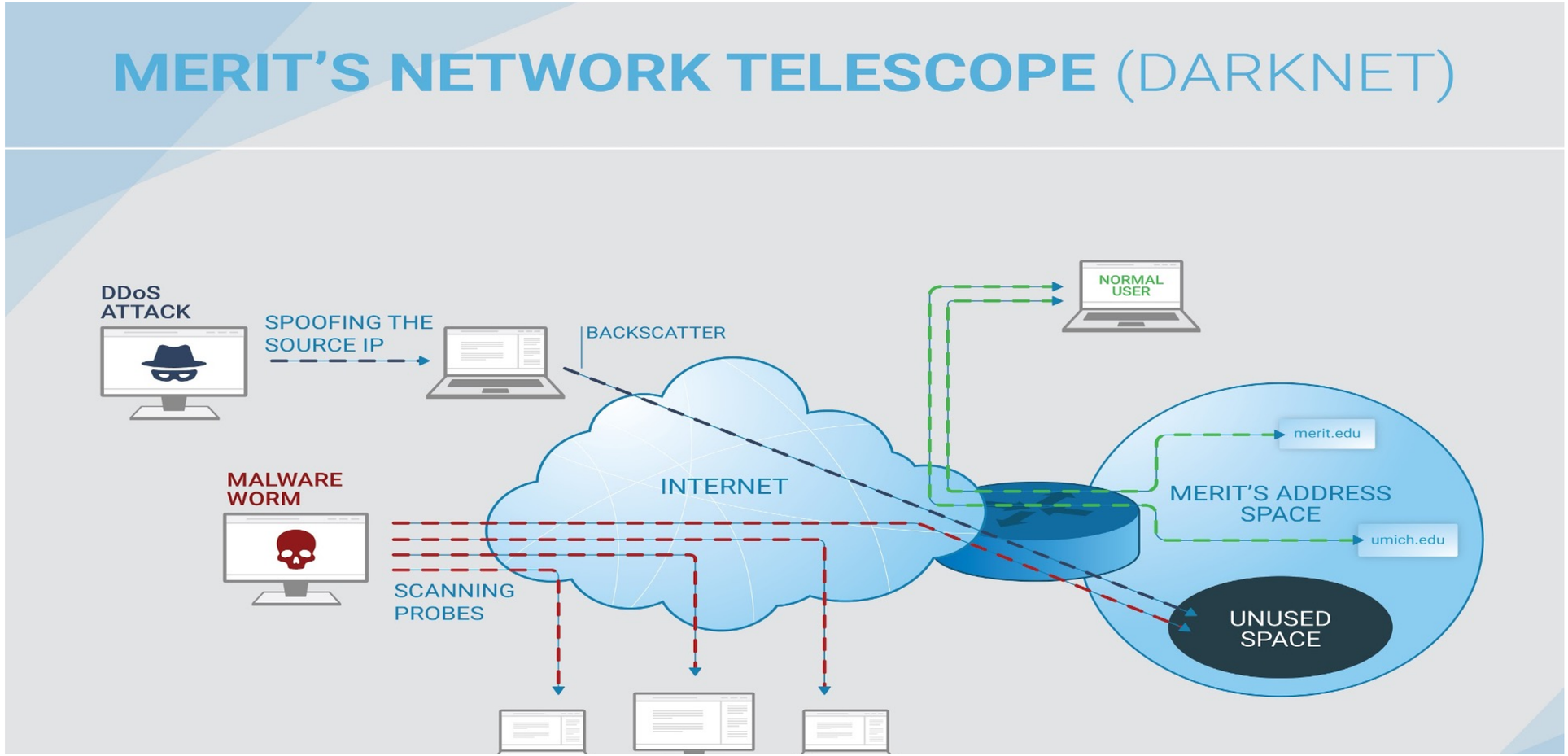
CAIDA DUST Workshop, July 2021

Acknowledgements

- NSF: CRI ORION: Observatory for Cyber-Risk Insights and Outages of Networks
    - Michalis Kallitsis, Zakir Durumeric (Stanford)
- NSF: ATD: Extremal Dependence and Change-point Detection Methods for High-dimensional Data Streams with Applications to Network Cybersecurity
    - Stilian Stoev (UM), George Michailidis (Florida), Michalis Kallitsis
- DHS S&T: CAOE: Characterizing Malware Behaviors using Darknet Data
    - John Yen (Penn State), Michalis Kallitsis

Talk Outline

1. Current developments with Merit’s network telescope
2. Reactive and distributed honeynet: early explorations
3. Research case study: classification of Darknet events



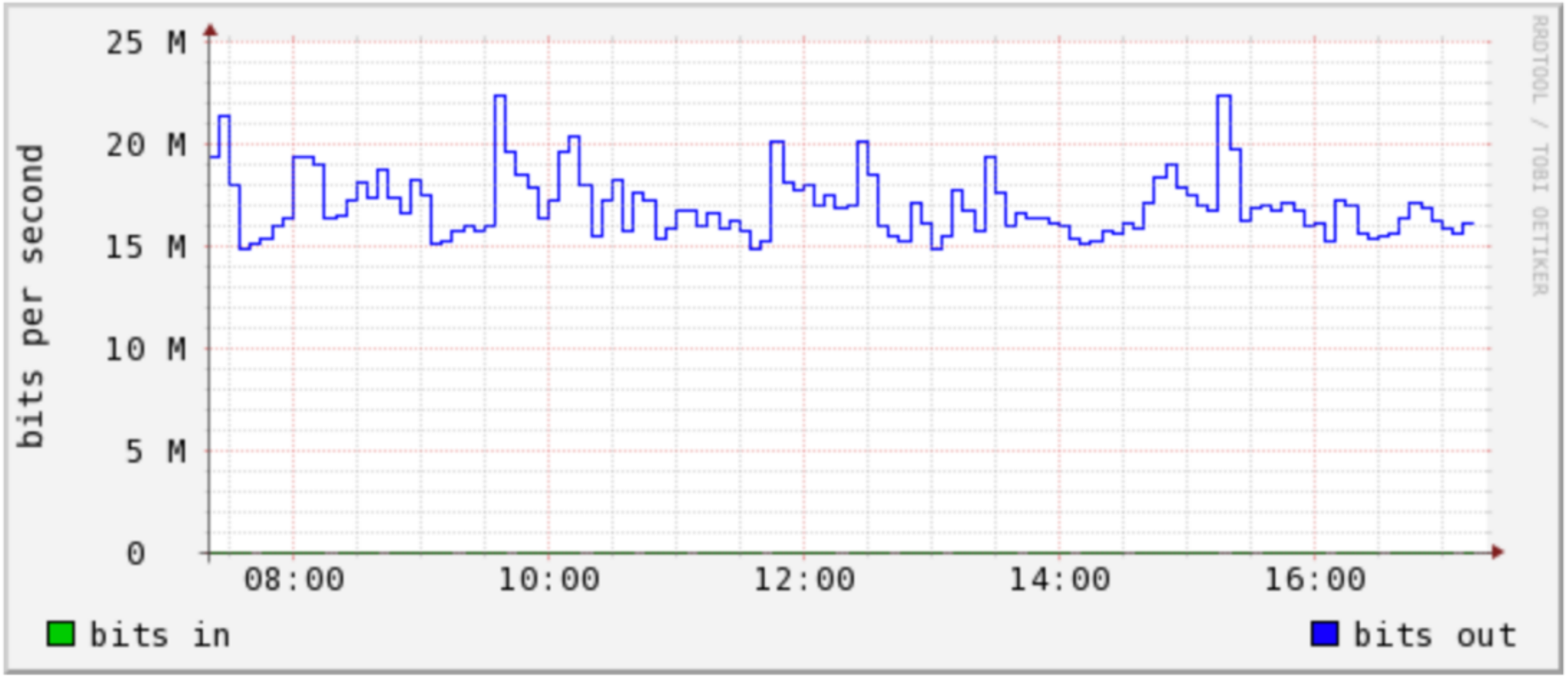
1. ORION network telescope

ORION network telescope in numbers

- Currently, approximately a **/13 subnet** (ie, about 500,000 unique IPs)
  - Down from our own /8 but still quite large
- 120GB/day compressed PCAP data
- Started renewing our infrastructure in 2018 with support from **NSF CRI** grant

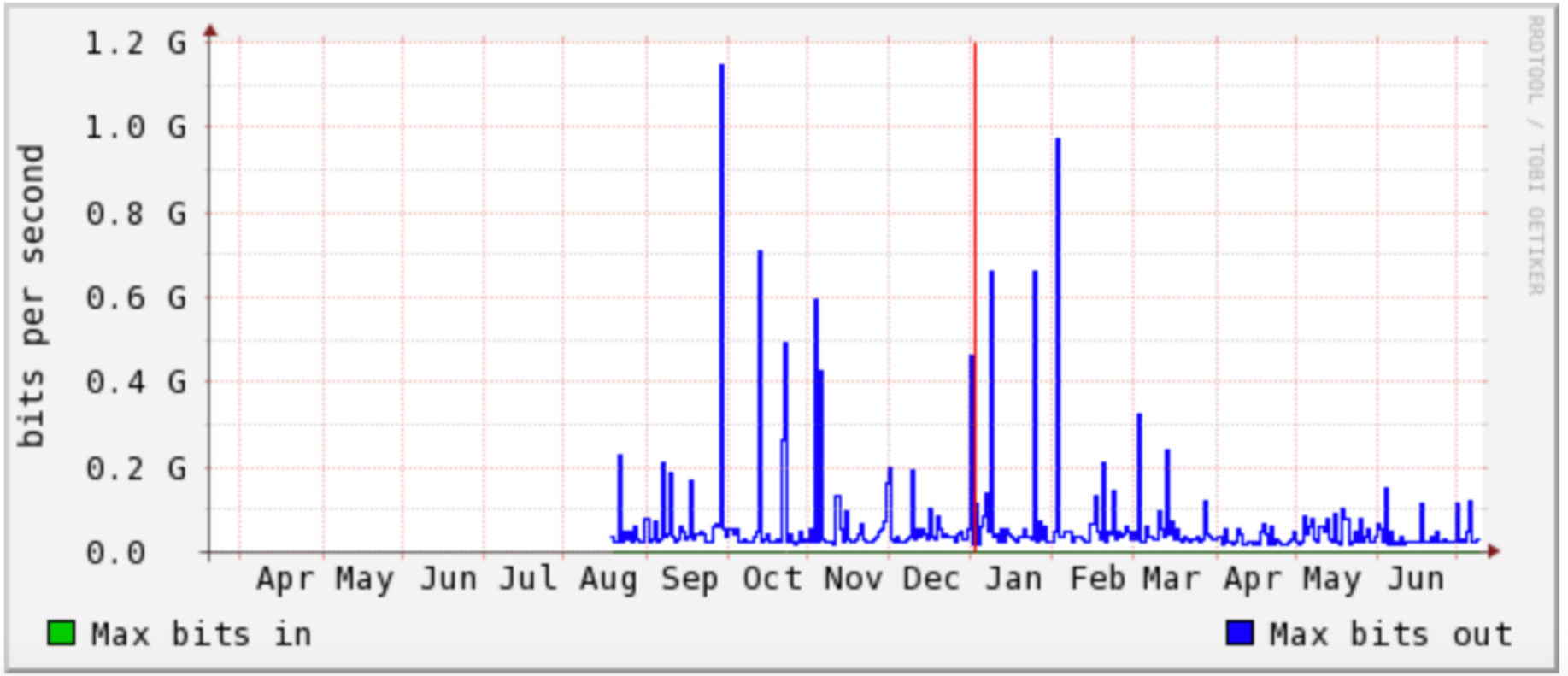
ORION network telescope in numbers — Darknet traffic

Hourly graph

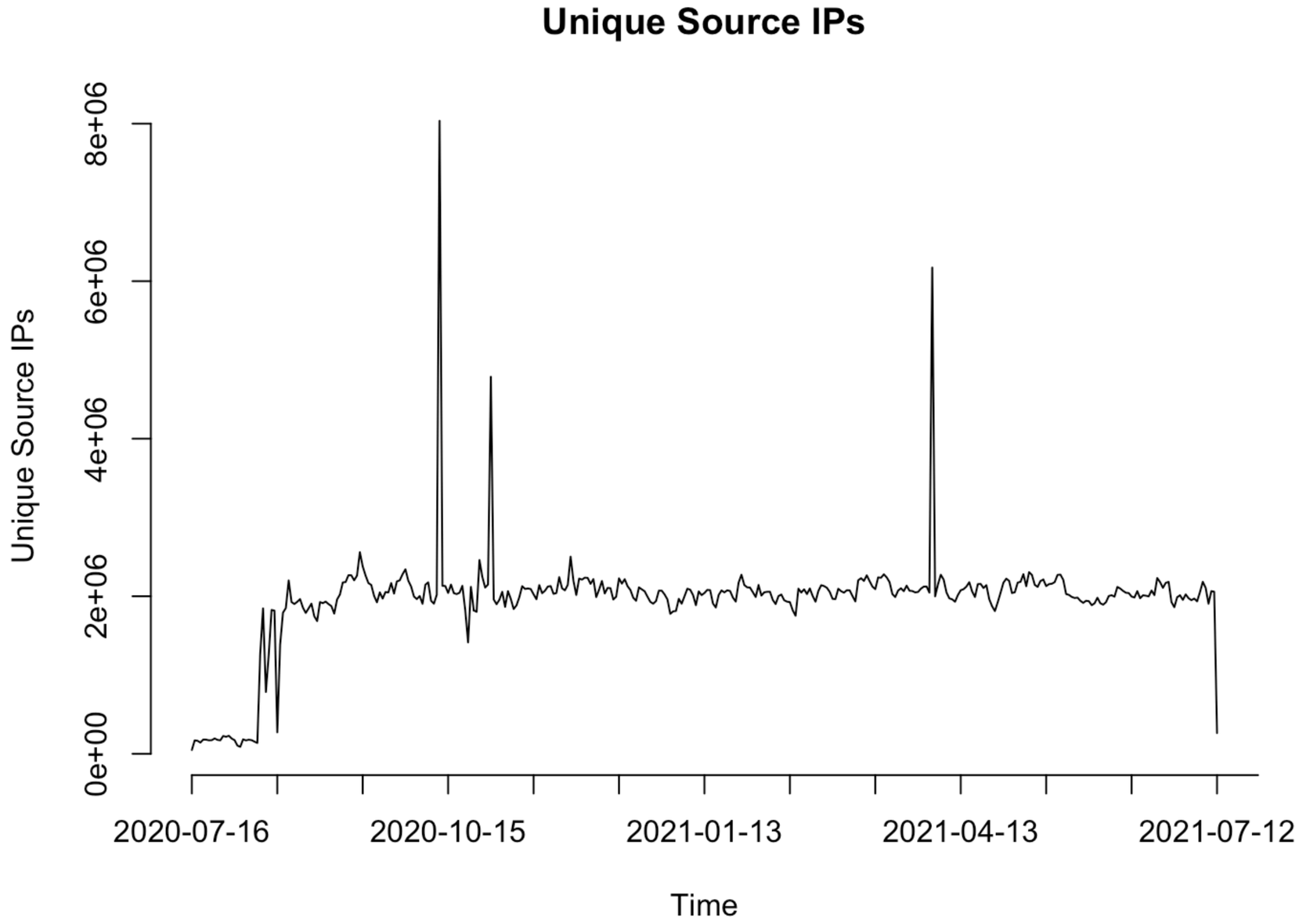


ORION network telescope in numbers — Darknet traffic

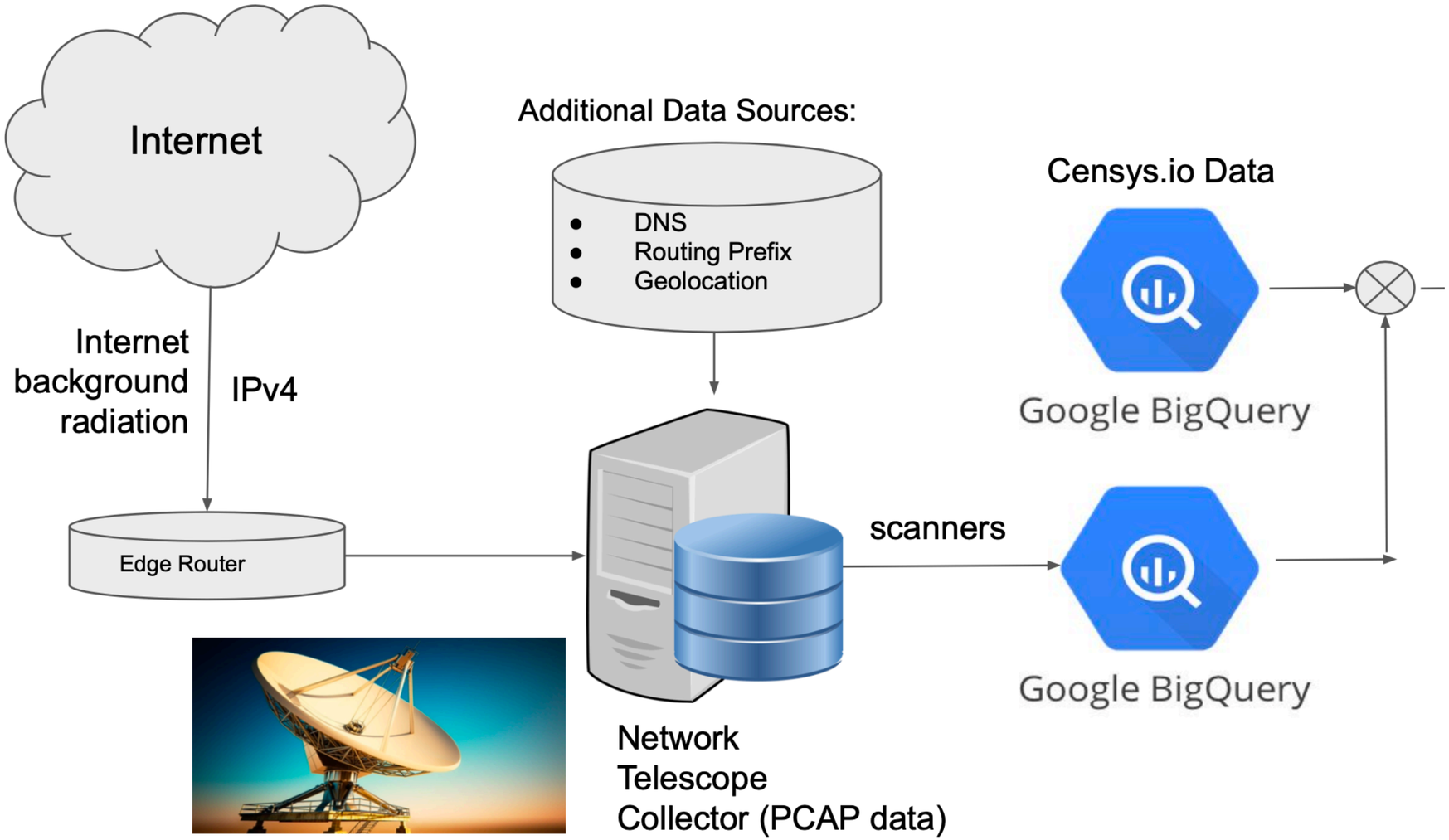
Yearly graph



ORION network telescope in numbers



ORION data pipeline



ORION data pipeline

- **Continuous** (i.e., state preserving) parsing of PCAP files
  - Parse on hourly basis
  - Extract “Darknet events”: Go software: https://github.com/Merit-Research/darknet-events
- Events “keyed” by (**source IP, port, traffic type**)
- **Traffic type** examples (full list: https://github.com/Merit-Research/darknet-events/blob/master/README.md):
  - TCP SYN (i.e., scanning)
  - ICMP Echo Request (i.e., scanning)
  - TCP SYN/ACK (i.e., backscatter)
  - TCP RST (i.e., backscatter)

ORION data pipeline — Config parameters

- **Timeout interval**: after how long to “expire” events and remove from cache
  - See “flow timeout problem”: Network Telescopes: Technical Report, Moore at al., https://www.caida.org/catalog/papers/2004_tr_2004_04/tr-2004-04.pdf
  - Typical longest gap “rule”: we use about 10 minutes (this would prevent “splitting” a scan with duration 2 days and rate 100pps for our Darknet size)
  - The Longest Run of Heads, Mark F. Schilling, https://www.jstor.org/stable/2686886
- **Unique destinations**: we just use 1
- **Samples**: store up to 3 packets using reservoir sampling

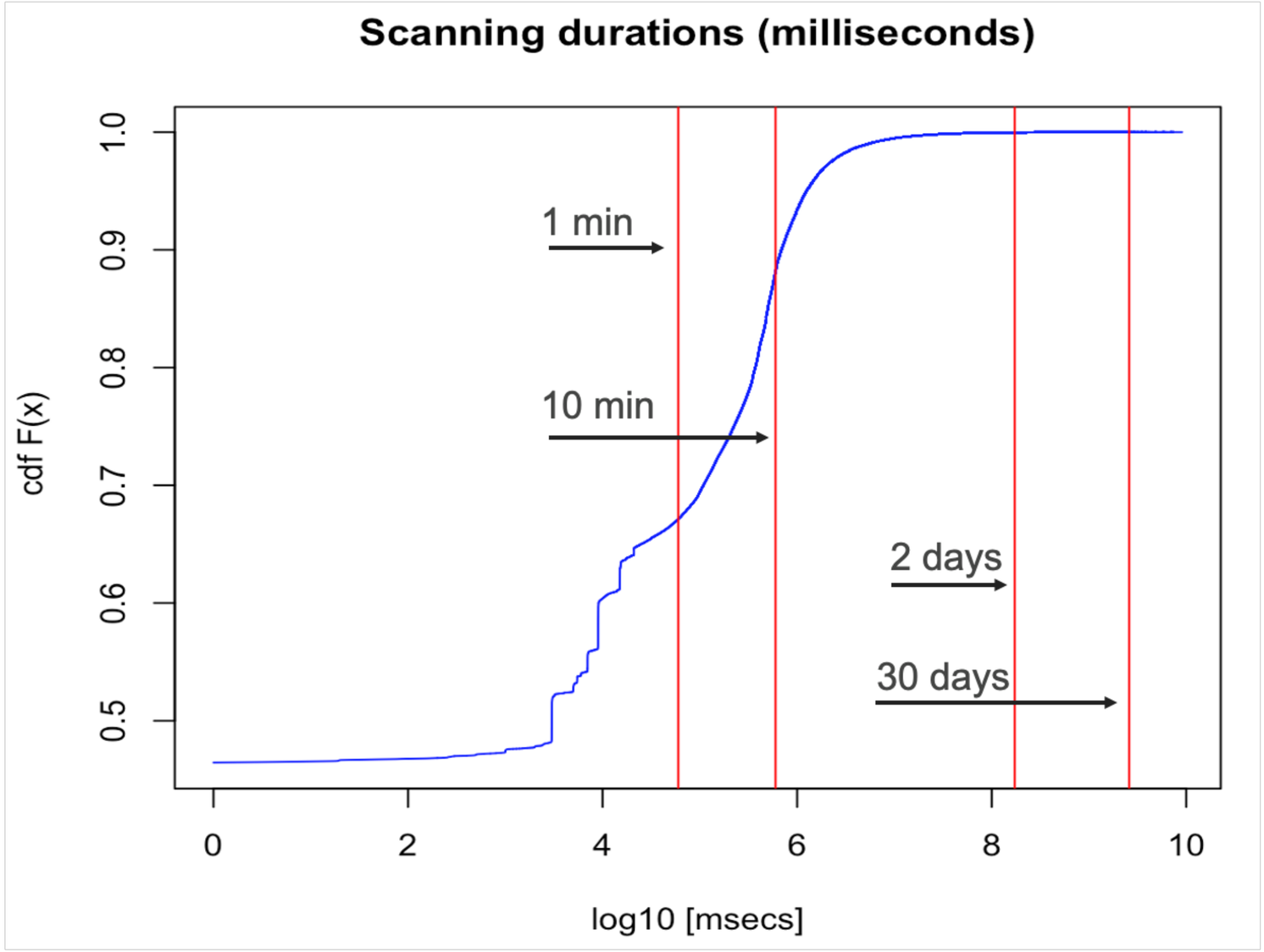
Approach Advantages

- Ease of data **analysis**: use standard SQL to process TBs in secs
- Ease of data **sharing**
  - Important: external users share the cost, i.e., “pay on demand” model
- Ease of data “**joins**” with external datasets (e.g., Censys, M-Lab)
- Ease of data **visualizations** for quick exploration (i.e., via Data Studio)
- Lossy **compression**

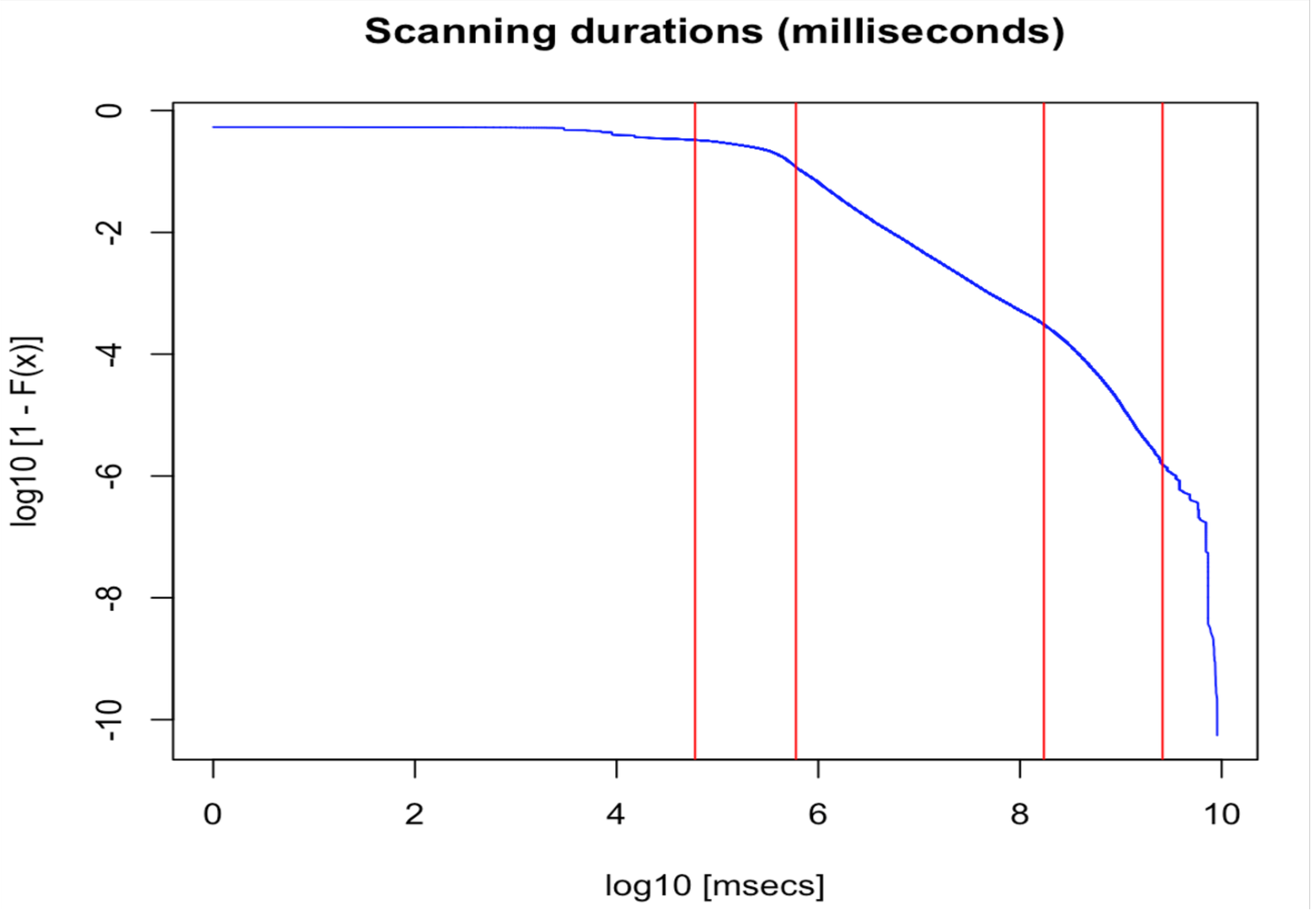
Approach Disadvantages

- **Lossy** compression
- Diminished ability for fine-grained **time series**
  - Though we have some ideas to (approximately) fix this
- Can become **expensive**
  - Storage and processing (queries) are charged
  - $5 per 1TB of data processed
- Handling control to a **3rd party**

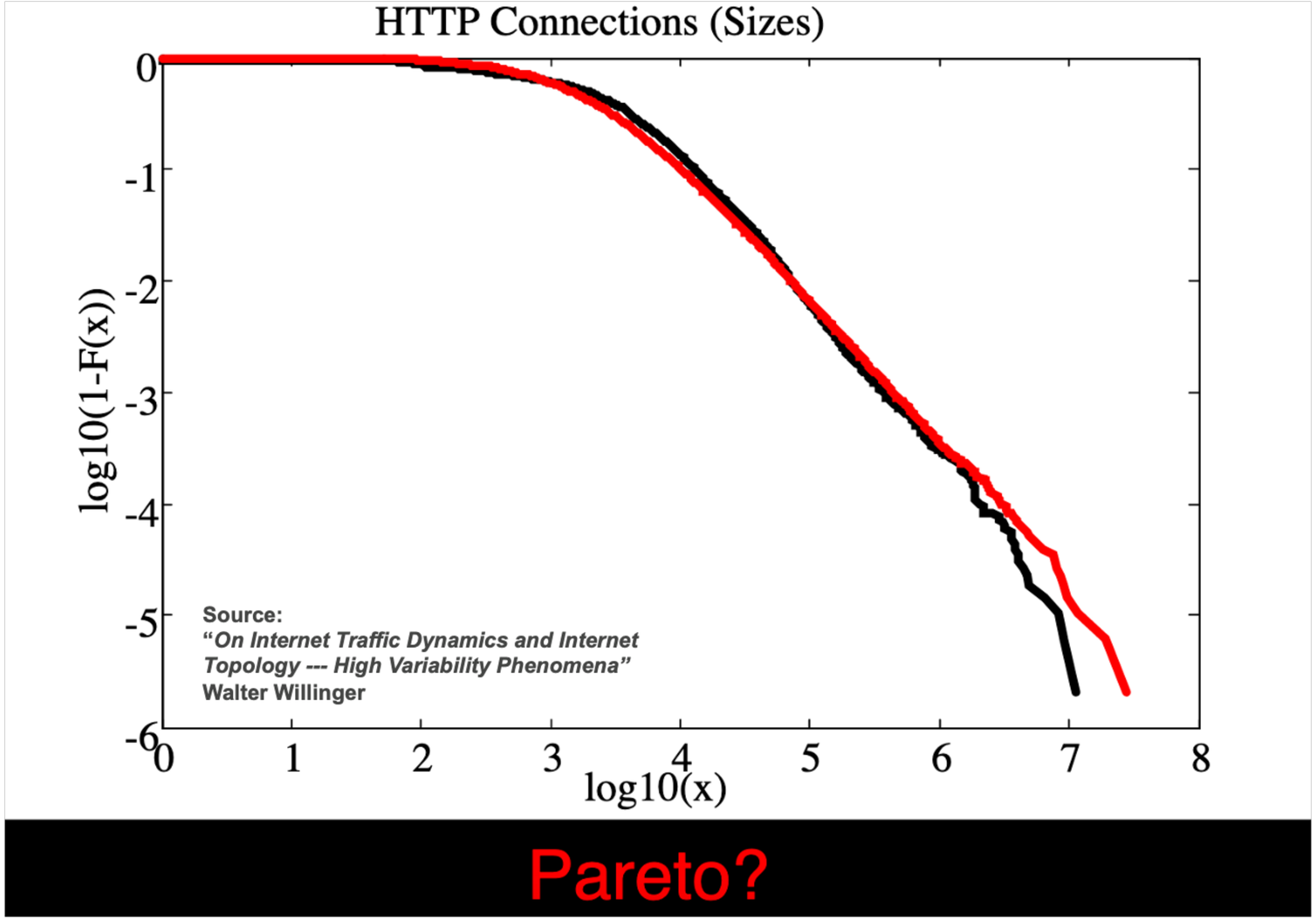
Analysis case study: scanning durations of 18 billion events



Analysis case study: scanning durations of 18 billion events



Analysis case study: scanning durations of 18 billion events

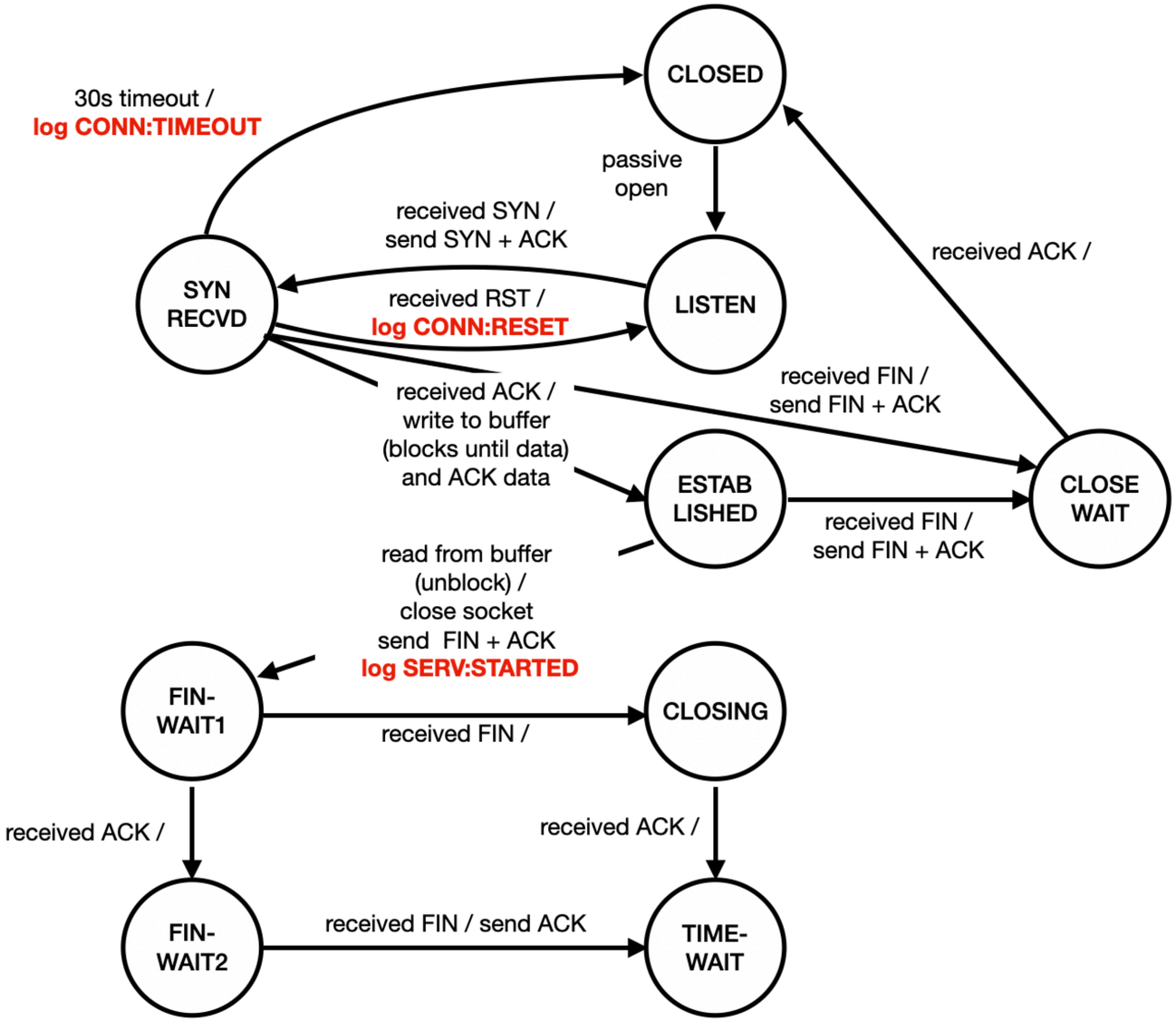


2. ORION Canary deployment

HoneyTrap Open Source project

- https://docs.honeytrap.io
  - Actively maintained by DTact (previously DutchSec)
  - Key contributor: Remco Verhoef
  - Offers 10 or so high-interactivity services (such as SSH, Telnet, etc.)
    - SSH simulator: Good for catching brute-force passwords and “executed commands”
- We instead use the “low interaction” Canary sensor within HoneyTrap
- **Why Canary?**
  - Monitor all ports (we sacrifice only one high-numbered TCP port used for SSH)
  - Get some extra visibility into the TCP payloads (when available)
  - Easy to distribute to multiple locations

Canary's TCP State Diagram



Canary deployment

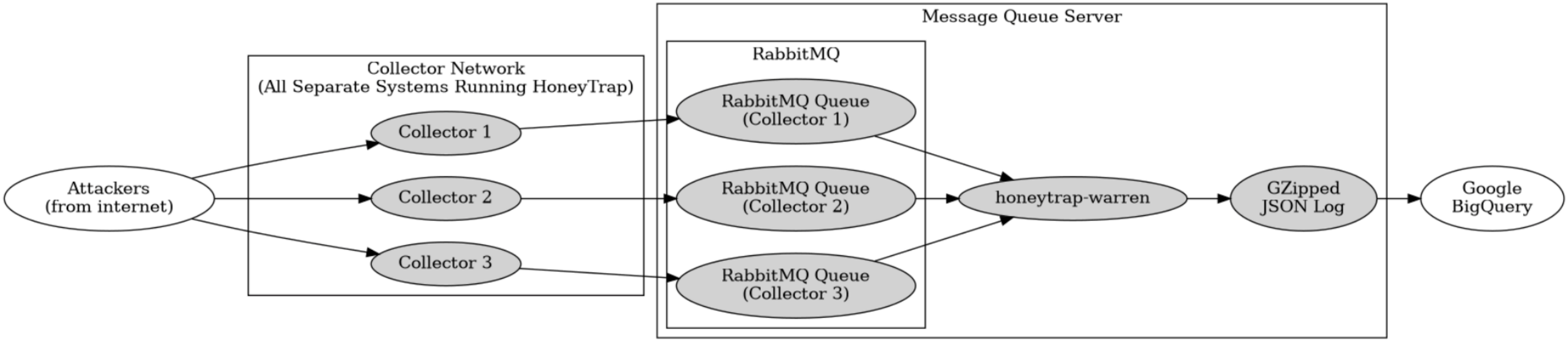


Figure 1: HoneyTrap Pipeline And System Architecture

Our Canary vantage points

- Academic institution 1: 64 unique IPs
- Academic institution 2: 1 IP
- Academic institution 3 (Merit): /24 deployment (currently 64 unique IPs)
- Cloud providers: AWS and Google (between 2 and 100 unique IPs)
- Orchestration managed with Ansible scripts which makes adding new nodes (relatively) easy

Summary

- New ORION infrastructure in production since 2020
- Data on BigQuery: easily sharable, analyzed, joined with other data
- Enables rapid experimentation / visualizations, supports our Darknet research (clustering, extremal dependence, etc.), supports data for education
- ORION's Canary: towards a distributed & reactive honeypot
    - Data also in BigQuery

Thank You!

mgkallit@merit.edu