STARDUST

Sustainable Tools for Analysis and Research on Darknet UnSolicited Traffic

Alistair King alistair@caida.org
Alberto Dainotti alberto@caida.org

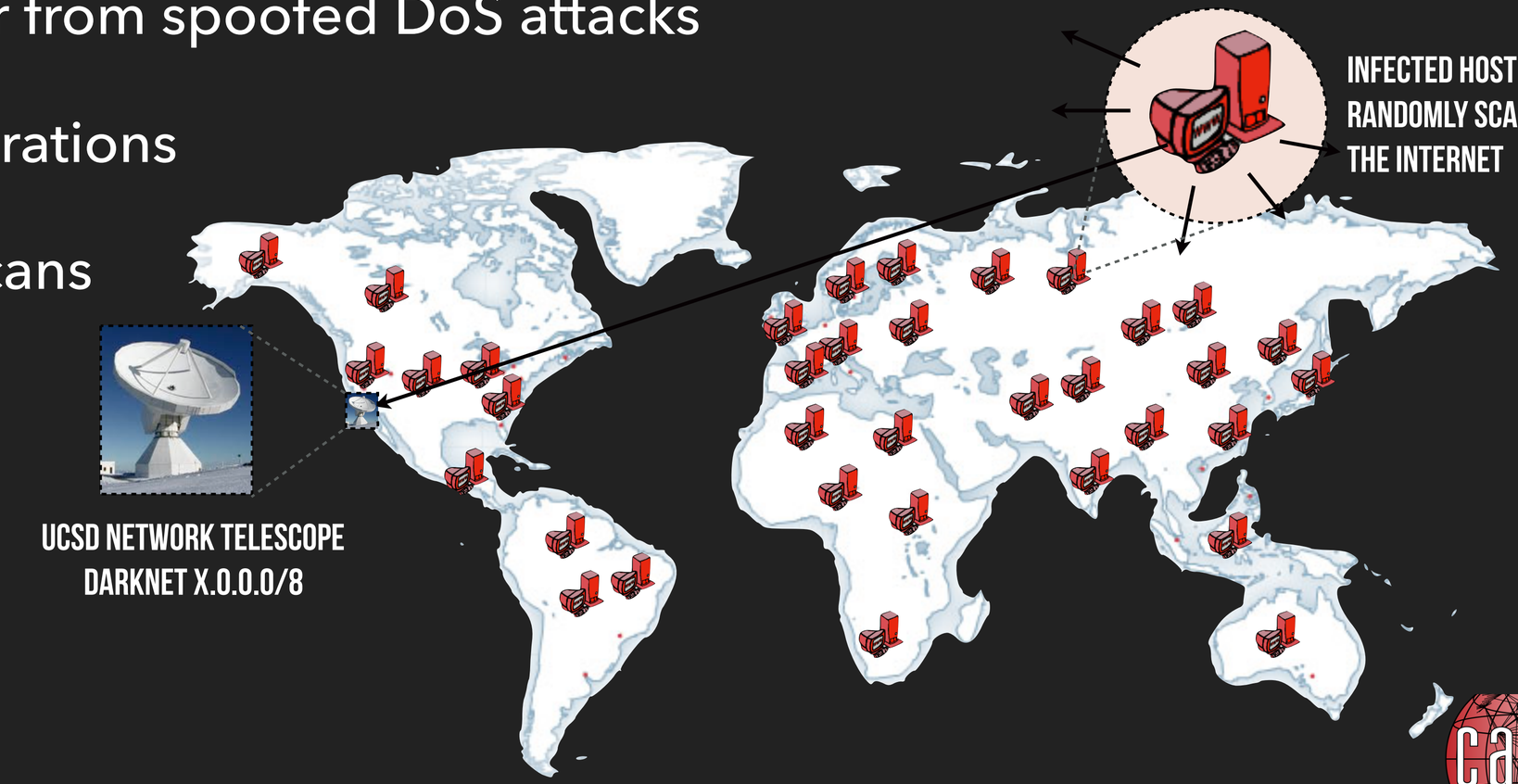
THE UCSD NETWORK TELESCOPE

- Passive traffic monitoring system
- Globally routed, lightly used /8 network
  (1/256 of the entire IPv4 address space)
- 24/7 full packet traces
- Archive of pcap data back to 2003
  (sampled data prior to 2008)
  - ~2 PB currently, growing by ~30 TB per month

HOW DOES IT WORK?

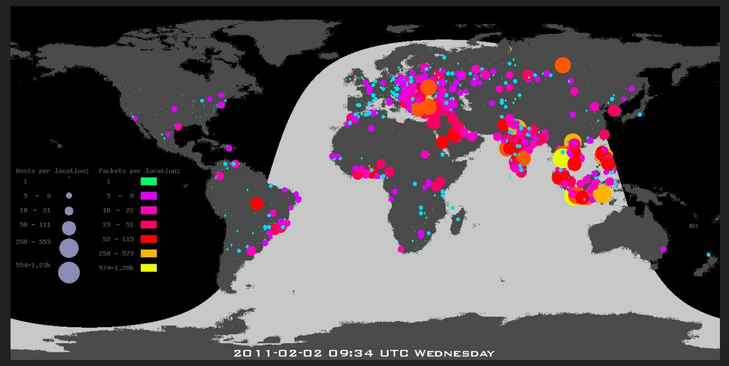
Who would send traffic to an unused network?

- Malware attempting to propagate
- Backscatter from spoofed DoS attacks
- Misconfigurations
- Network scans
- …

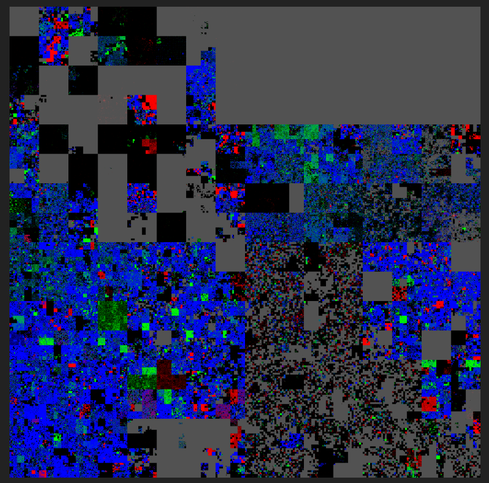


WHAT IS IT GOOD FOR?

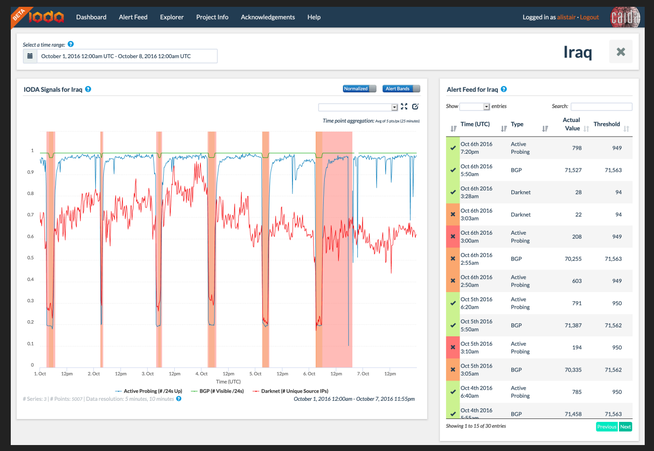
What can this data be used to study?



Malware Phenomena



IPv4 address space usage



Connectivity Disruptions

… and much more

(more than 100 scientific publications and PhD theses without CAIDA co-authorship)

USING THE DATA

- Raw pcap data has PII, so treated as sensitive
  (e.g., IP addresses, UDP payload)
- Researcher access via code-to-the-data approach
  - Apply via DHS IMPACT portal (https://www.impactcybertrust.org/)
  - SSH access to shared CAIDA-operated UNIX machine
  - Recent pcap and flow-level data
    (historical pcap can be retrieved from tape archive)
- Single research-compute machine means resource contention
  - pcap files are huge (~200 GB/hour),
    analysis is complex & time consuming

STARDUST: TOWARD A MORE SUSTAINABLE FUTURE

- Three primary goals:

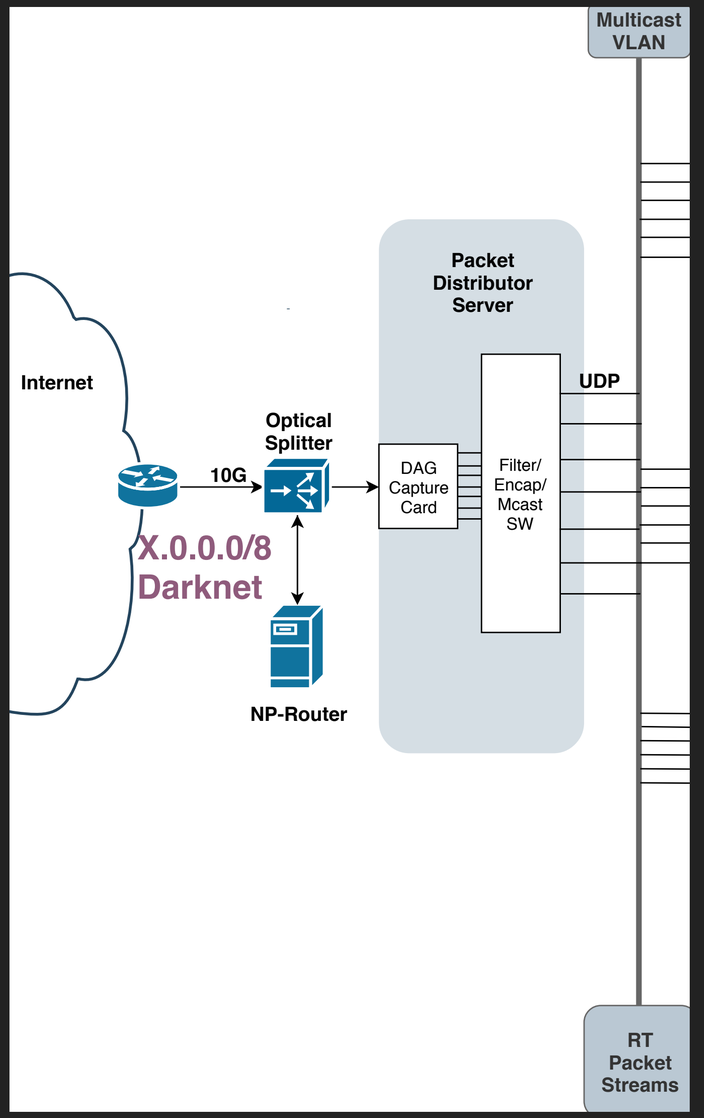
1. Scalable, future-ready traffic capture and real-time distribution system

2. Flexible, extensible, sustainable Research-Compute (RC) environment by leveraging modern virtualization and containerization technology (e.g., Kubernetes) as well as NSF-funded supercomputers (e.g., SDSC's Comet)

3. Lower the barrier to entry for new researchers, and reduce Time-To-Insight by providing high-level, annotated datasets

1. HIGH-SPEED CAPTURE AND REAL TIME DISTRIBUTION

- Endace 10 Gbps DAG card
- Multi-threaded packet distribution software
  - Captures from DAG card
  - Filters out “legitimate” traffic
  - Publishes packet batches to multicast group(s) on dedicated VLAN
  - Configurable routing of packets to streams (e.g., send XX.YY.0.0/16 to “small darknet” stream)
- Clients connected to VLAN can process packets from a stream using libtrace API or tools
- Developed in collaboration with WAND group at Waikato NZ



2. SCALABLE RESEARCH-COMPUTE (RC) ENVIRONMENT

- Containerization, of course
  - Decouple compute from hardware
  - Customized environment (scripts etc.)
  - Portable (e.g., move heavy users to supercomputer)
- Real-time traffic feeds (VLAN) available in container
- Historical data in object storage cluster (>500 TB, 60 Gbps)

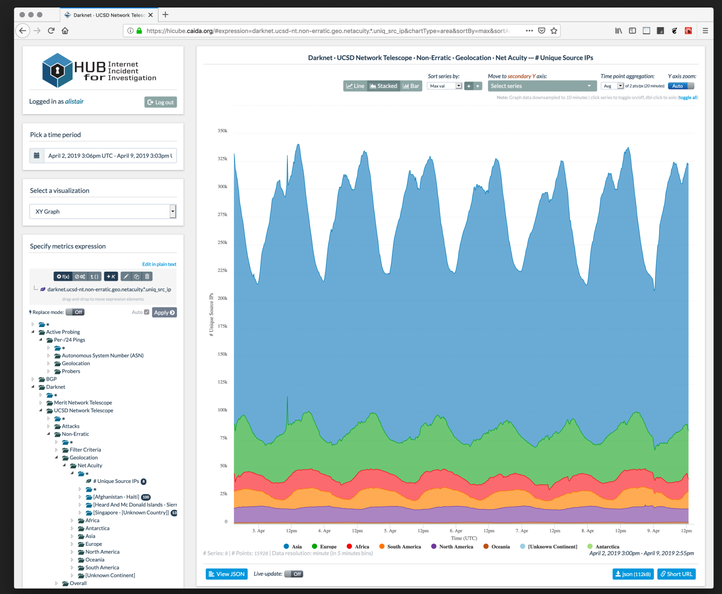
- Current proof-of-concept deployment with VMs

3. HIGH-LEVEL ANNOTATED DATASETS

- Processing raw pcap is hard (even for experts)
  - Huge volume (~1 Mpps, ~200 GB/hr)
  - Longitudinal processing virtually impossible (> 600 TB/yr)
- Post-processed, aggregated, annotated datasets
  - Lower barrier to entry for new researchers
  - Reduce time to insight
  - Facilitate “Big Data” analysis
- Files generated in near-realtime (< 1 hr latency)

3. HIGH-LEVEL ANNOTATED DATASETS (CONT.)

- Annotated “Flow”-level data
  - Src-IP, Dst-IP, Src-Port, Dst-Port, Proto, TTL, TCP-Flags, Len
    Country, Region, Lat/Long, ASN, Is-Spoofed
- Inferred Randomly Spoofed DoS attacks
- Time Series
  - Per-country, region, ASN, port, protocol, etc.
  - Used by IODA (outage detection),
    HI3 (cybersecurity event analysis)
- Supports longitudinal analysis currently,
  real-time streams coming (2020)



USE CASES

1. Classroom/Lab
   - Create customized container with scripts etc.
   - Each student/team uses one container
   - Exercises can target processing raw pcap or flow-level to find events
   - … or even real-time detection

USE CASES (CONT.)

2. Study scanning over time
   - One-off longitudinal analysis
   - Process full history of Flow data on dynamically provisioned Spark cluster
   - Identify groups of records indicative of scanning
3. Detect spoofing as it happens
   - Continuous real-time monitoring
   - Run in dedicated RC container
   - Execute active measurements in response

CURRENT STATUS

- New capture/distribution running since July 2018
- Experimental VM-based RC environment
  - VMs can trivially attach to live stream
- Experimental active anti-spoofing approach (consuming live stream from VM)
- All existing data moved to object storage cluster
  Big Data analytics at > 10Gbps

NEXT STEPS: SOFTWARE / INFRASTRUCTURE

- First VM-based beta users (Spring/Summer)
- Active anti-spoofing platform (Summer)
- Big Data analysis environment (Fall)
- Experiment with containerized RC environment (Fall)

- Data from additional telescopes coming soon:
  - Merit Networks
  - Politecnico di Torino, Italy
  - UFMG, Brazil

NEXT STEPS: WORKSHOPS

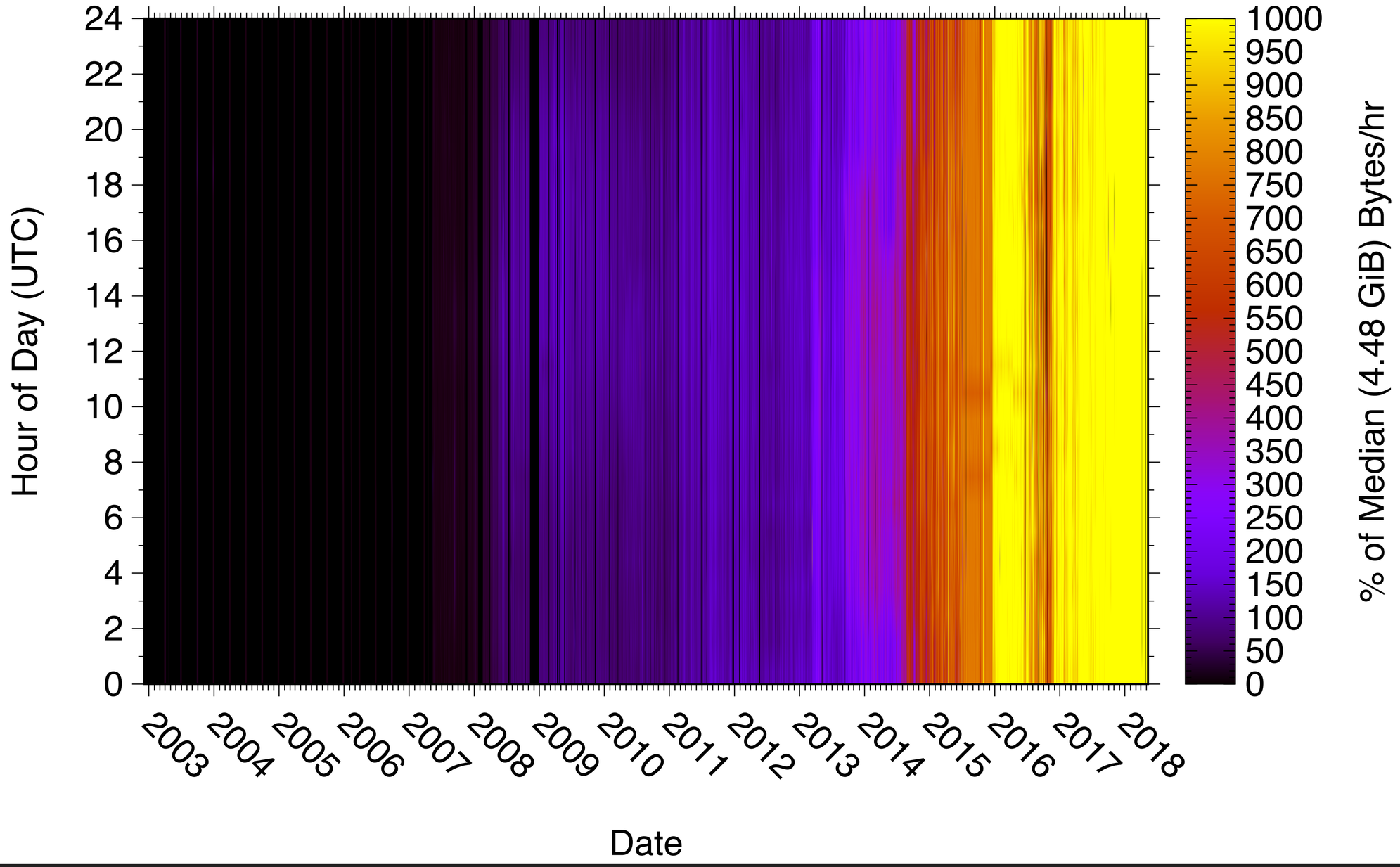
- 2019 DUST Workshop
  - September 9-10
  - Community building
  - Identifying research questions
  - Influence platform development
- 2020 DUST Workshop
  - Open platform to users
  - Hackathon

QUESTIONS?

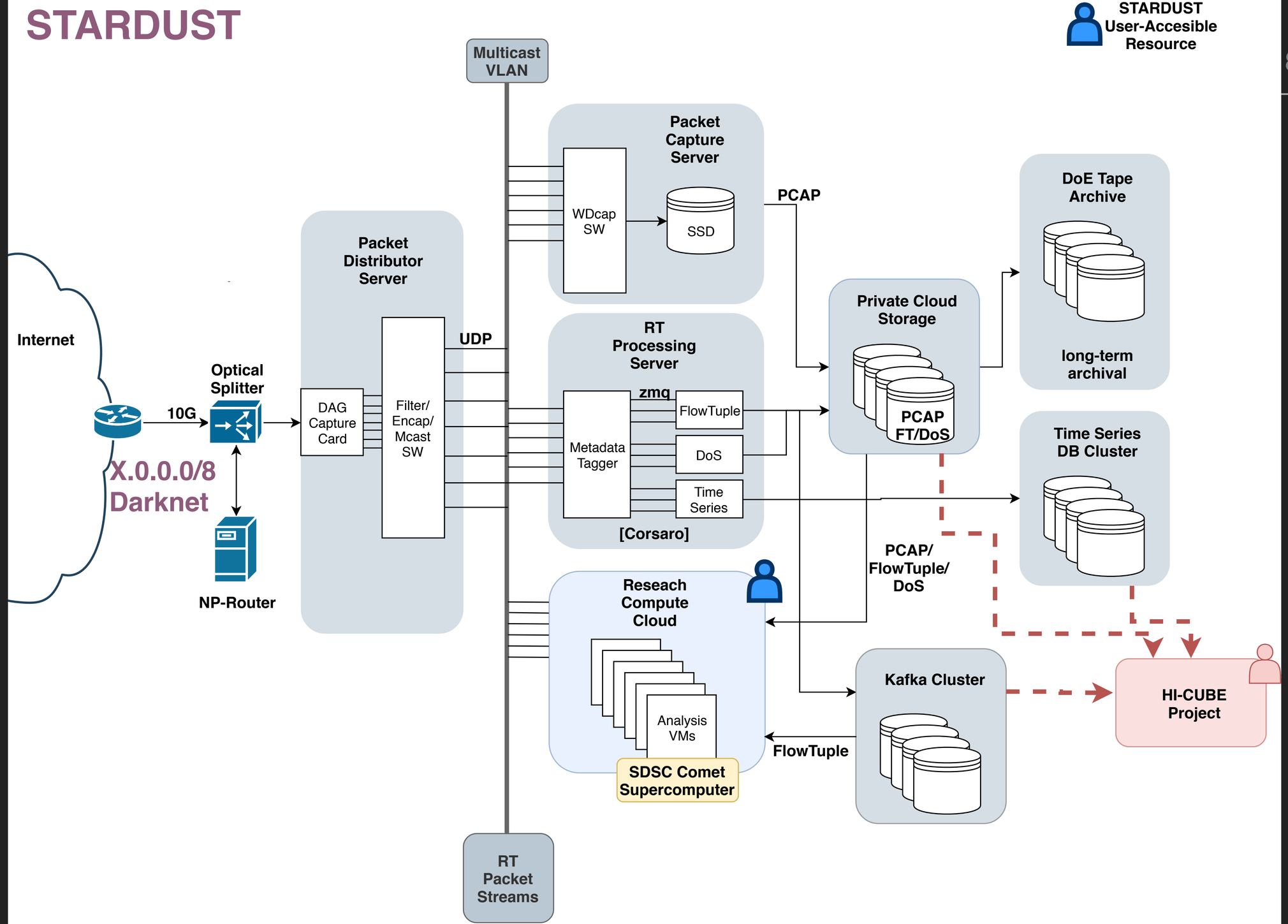
alistair@caida.org

caida.org/funding/stardust/

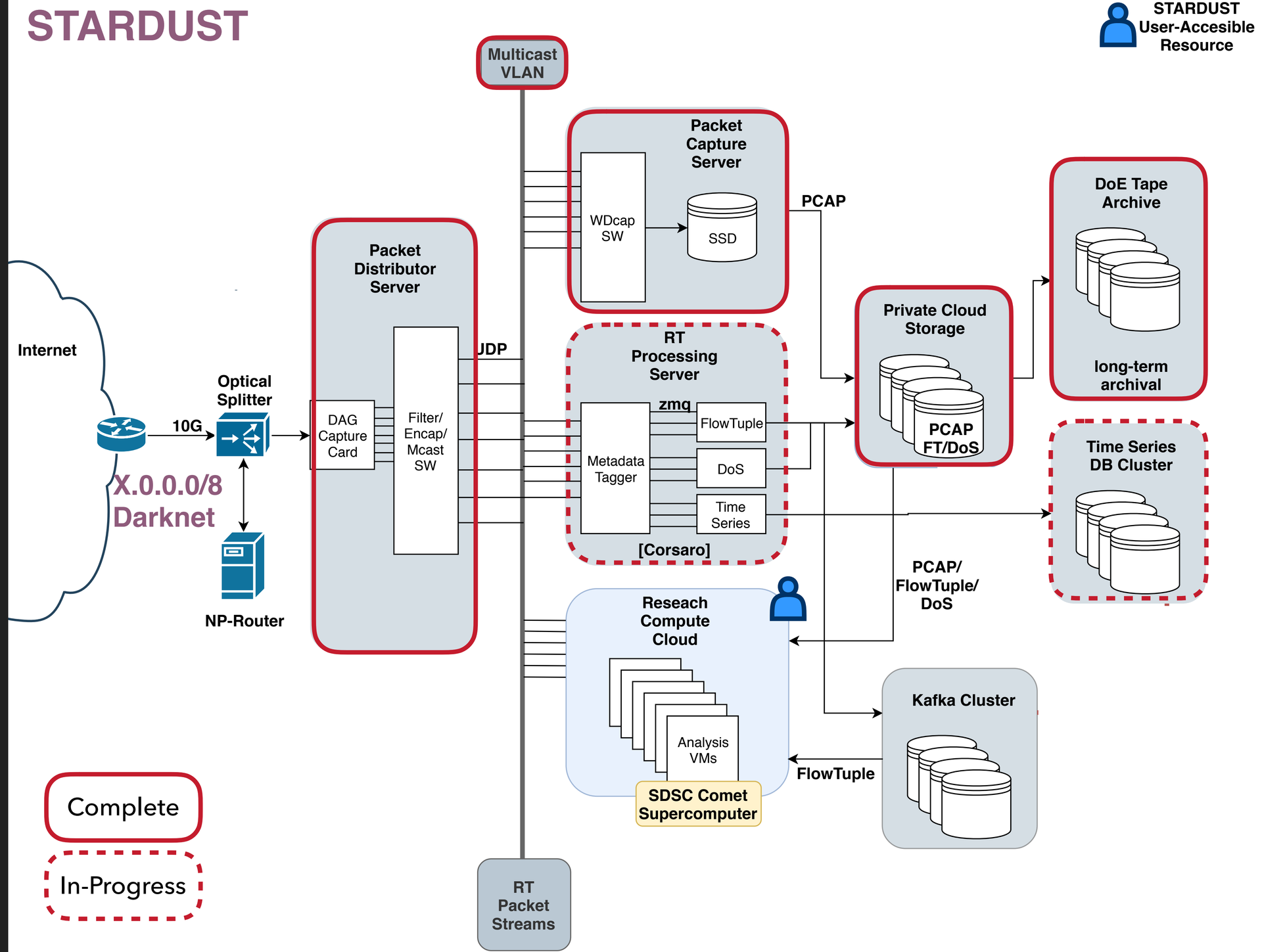
UCSD Network Telescope
Hourly Compressed Pcap Sizes



STARDUST


STARDUST
User-Accesible Resource
Multicast VLAN
Packet Capture Server
WDcap SW
SSD
PCAP
DoE Tape Archive
long-term archival
Private Cloud Storage
PCAP FT/DoS
Packet Distributor Server
UDP
RT Processing Server
zmq
FlowTuple
DoS
Time Series
Metadata Tagger
[Corsaro]
Internet
Optical Splitter
10G
DAG Capture Card
Filter/ Encap/ Mcast SW
X.0.0.0/8 Darknet
NP-Router
Time Series DB Cluster
PCAP/ FlowTuple/ DoS
Reseach Compute Cloud
Analysis VMs
SDSC Comet Supercomputer
Kafka Cluster
FlowTuple
HI-CUBE Project
RT Packet Streams

STARDUST


STARDUST
User-Accesible Resource
Multicast VLAN
Packet Capture Server
WDcap SW
SSD
PCAP
DoE Tape Archive
long-term archival
Private Cloud Storage
PCAP FT/DoS
Packet Distributor Server
DAG Capture Card
Filter/ Encap/ Mcast SW
UDP
RT Processing Server
zmq
Metadata Tagger
FlowTuple
DoS
Time Series
[Corsaro]
Time Series DB Cluster
Internet
Optical Splitter
10G
X.0.0.0/8 Darknet
NP-Router
PCAP/ FlowTuple/ DoS
Reseach Compute Cloud
Analysis VMs
SDSC Comet Supercomputer
Kafka Cluster
FlowTuple
RT Packet Streams
Complete
In-Progress