DHS PREDICT project: CAIDA update

Kimberly Claffy, CAIDA

SRI Rosslyn, Washington D.C.

5 November 2012

DHS PREDICT project: CAIDA update

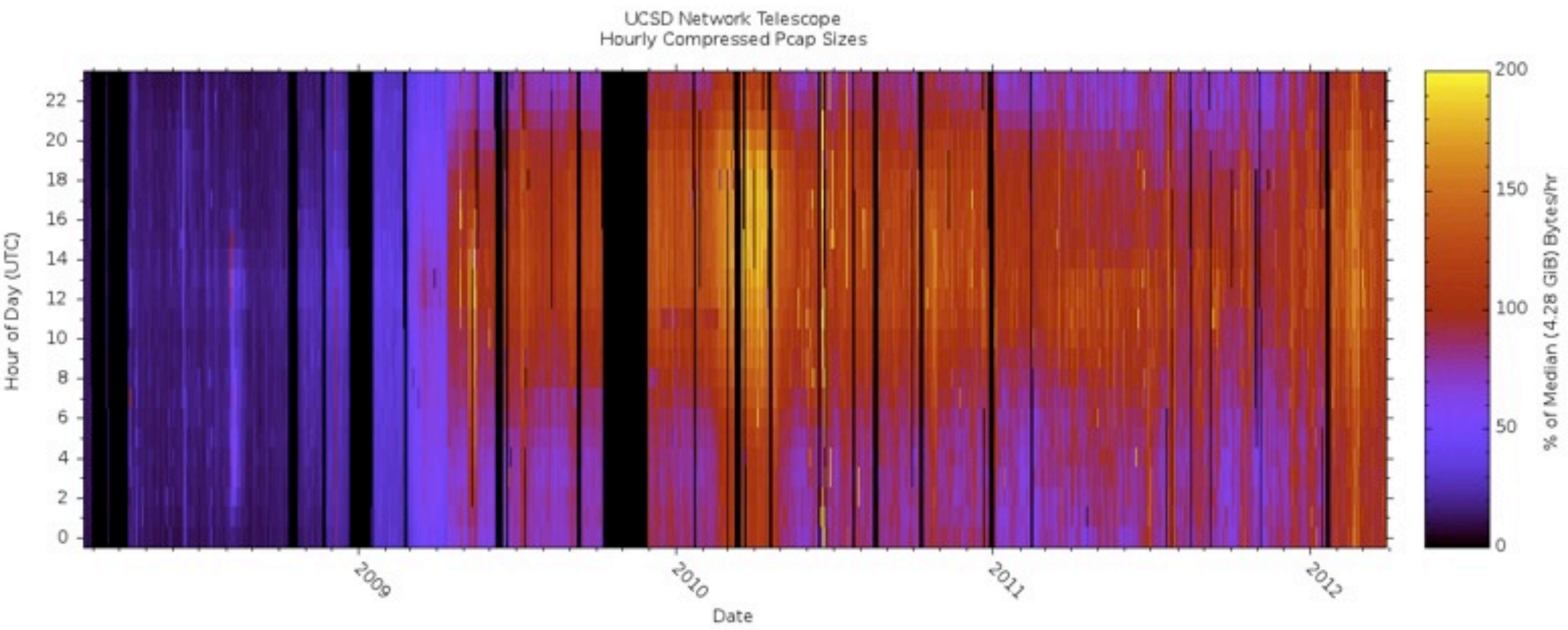
- Data storage status
- Data collection status
- Data set dissemination statistics
- Other activities
- Open issues

Data Storage Status

- Real-time telescope data (last 60 days, plus selected periods now stored on dedicated CAIDA-owned data server (thor.caida.org) with 53 TB of RAID storage.
- raw telescope data (~150TB) archived on HPSS tape at NERSC
- All other CAIDA data, incl. all online CAIDA datasets stored on two CAIDA-owned data servers (thoth and indy) w/a combined ~50 TB disk space (11TB free).
- Non-telescope data uses 20 TB SDSC allocation,, on the SDSC cloud storage system (http://cloud.sdsc.edu).
- Purchase more storage in the future as necessary.

Data Storage Status

- Transferred 10+ TiB to DOE lab NERS
  - National Energy Research Scientific Computing Center
  - One week to transfer 100TB on 22 March 2012
  - http://blog.caida.org/best_available_data/2012/04/04/targeted-serendipity-the-search-for-storage/



Data collection - passive

- High-speed backbone: March 2008 - Oct 2012
  - unanonymized: 10.6 TiB compressed, 21.2 TB uncompressed
  - anonymized: 9.5 TB compressed, 19.6 TB uncompressed
- Problems:
  - Chicago monitors offline since September 2011. Replacement hardware in Chicago, final testing blocked on remote hands.
- Status:
  - 2012 data sets online through October
  - took DITL trace, added to 2012 data set
  - Three traces around IPv6 Launch day (June) to be added to the online IPv6 Day dataset (includes traces from IPv6 Day in 2011).
  - Selection of 'best' quarterly traces complete up to third quarter in 2012. Currently the only consequence of this selection is that only these quarterly traces are backed up to the SDSC cloud system. Still serve all 12 months/year in annual online passive trace datasets.

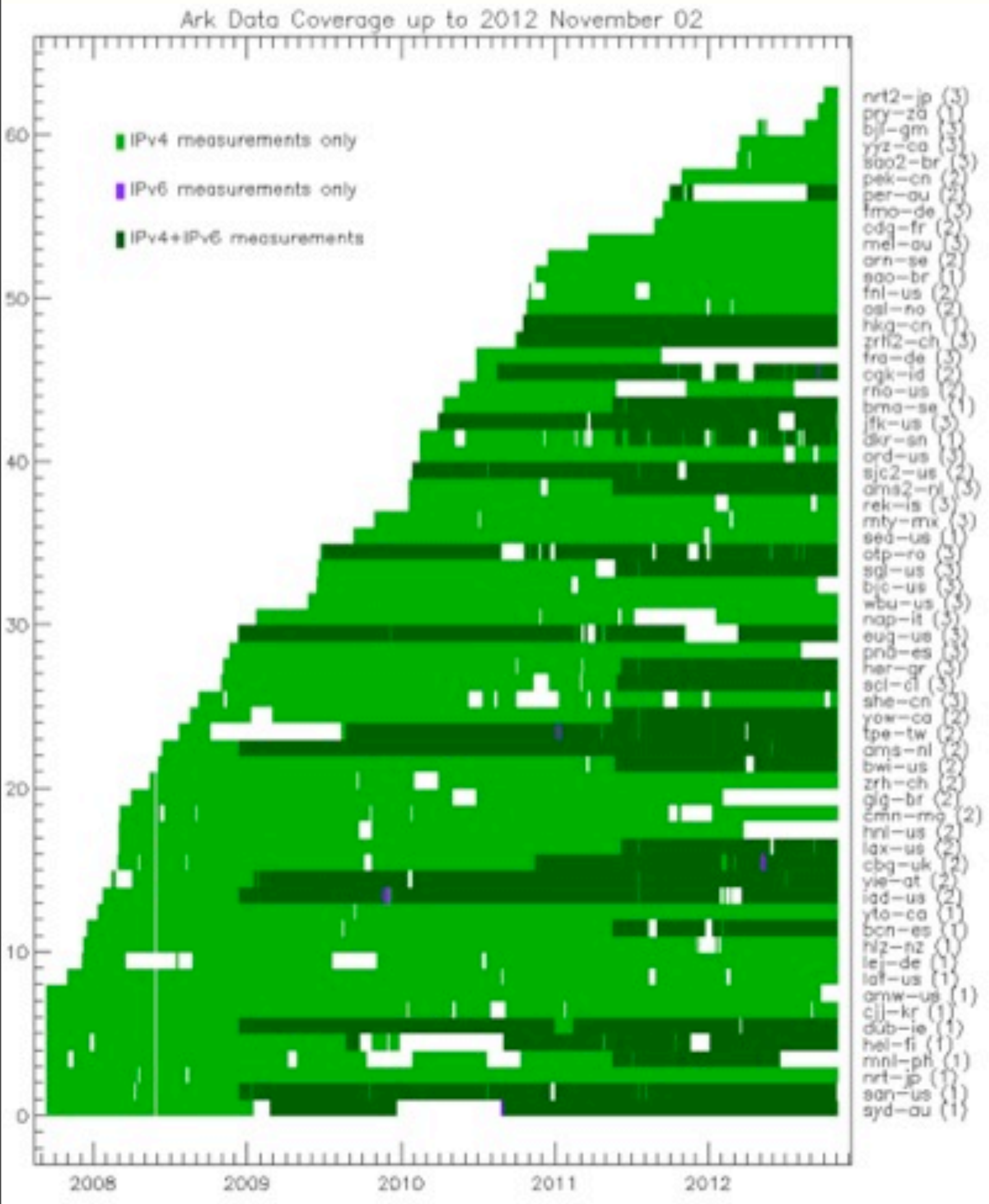
Data collection - passive

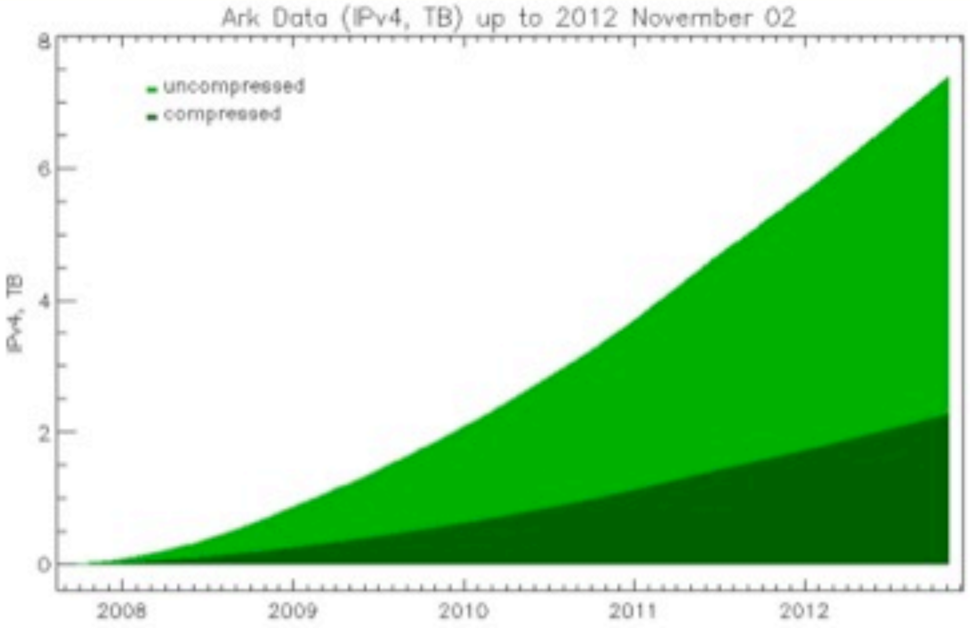
- UCSD telescope:
  - Most 2012 data still ‘live’ on disk (211 or 307 days so far)
    - 26.64 TiB compressed, 52.92 TiB uncompressed.
  - older data archived to NERSC in April
    - 134.38 TiB compressed/encrypted
  - summaries (Corsaro 8-tuples) stored on local disk, alongside raw data in active research use
- OC48 traces:
  - 964.5 GB (compressed), 1.7 TB (uncompressed)
  - unanonymized: 815.7 GB (compressed), 1.5 TB (uncompressed)
  - anonymized: 148.8 GB (compressed), 285.2 GB (uncompressed) (in PREDICT)

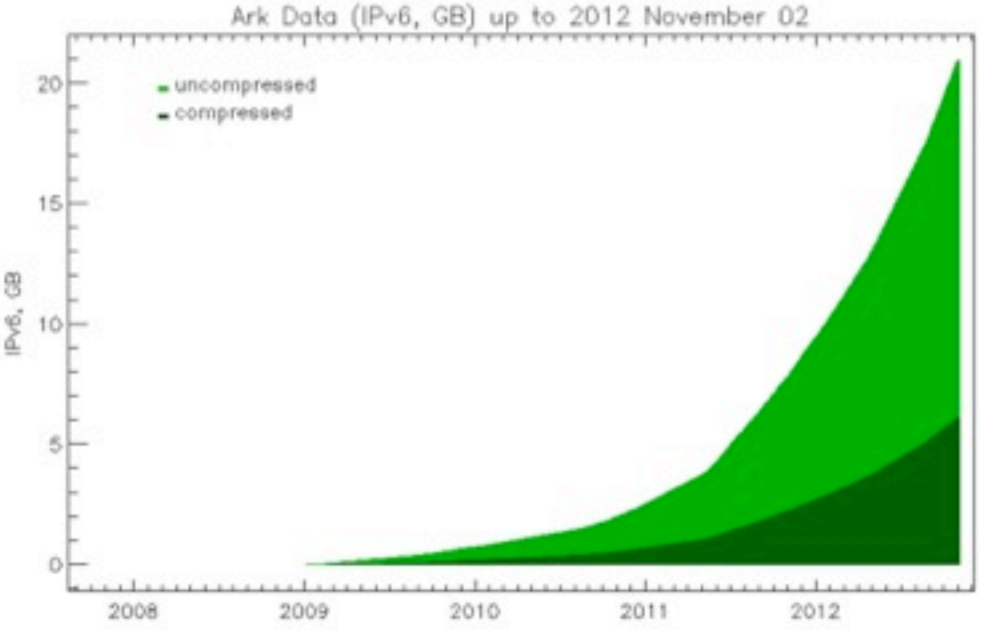
Data collection - active

- old skitter data (in PREDICT):
  - 1.47 TB (compressed), 4.02 TB (uncompressed)
    - discontinued February 2008
    - skitter ITDK now a public dataset
- current Ark data:
  - IPv4 topology: 2.3 TiB (compressed), 7.5 TiB (uncompressed)
  - IPv6 topology: 6.1 GiB (compressed), 21 GiB (uncompressed)
  - 62 monitors (and growing) in 32 countries, 28 IPv6 capable
- data curation:
  - create derivative data sets
  - aggregate in http://www.caida.org/data/active/internet-topology-data-kit/
    - last ITDK added July 2012
    - router-level topologies: nodes and links
    - host names, AS names, geographical info, AS relationships

Archipelago Monitors and Data





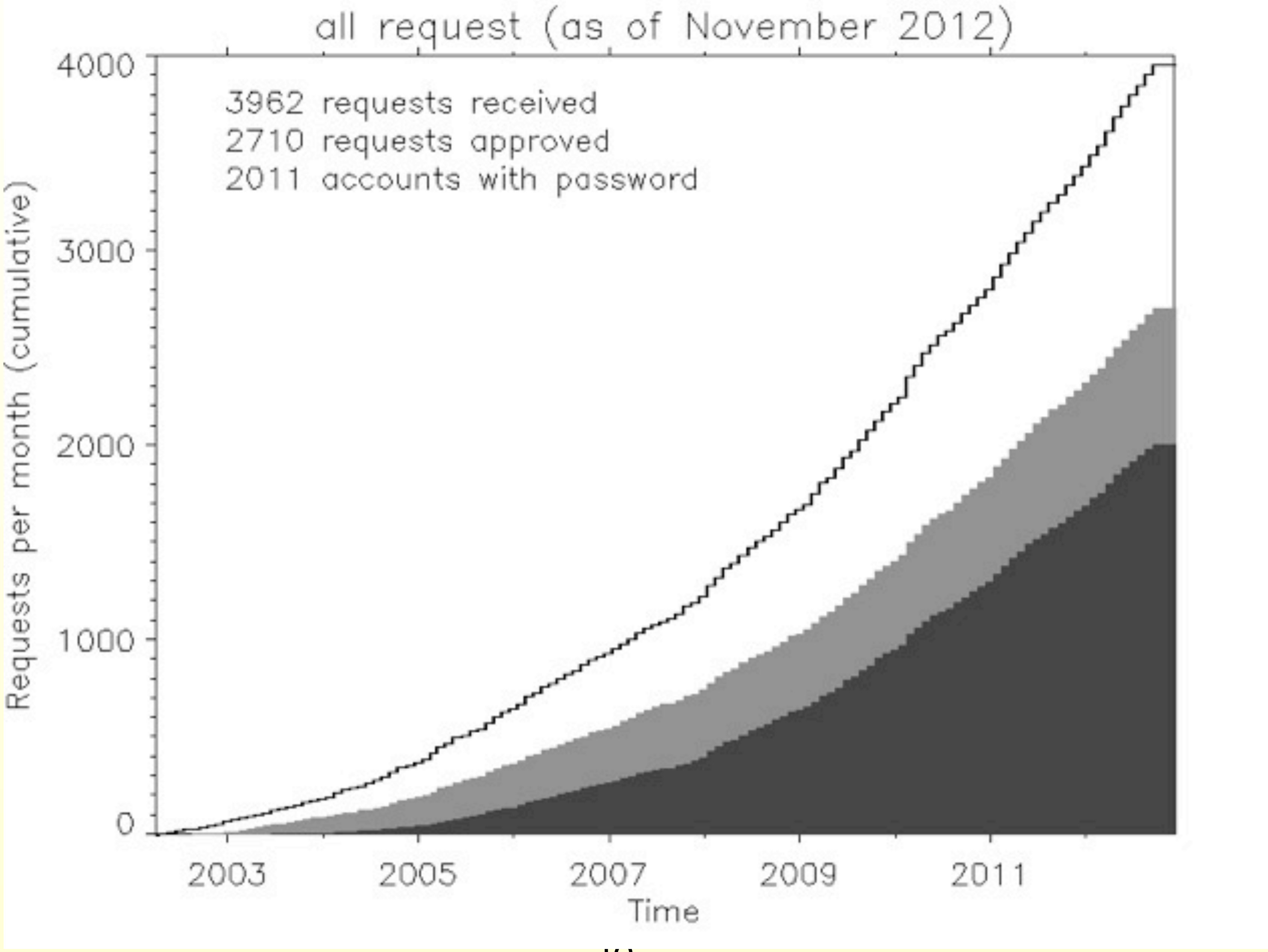


Requests for the data, 2012/2011/2010/2009

Dataset	Requests	Approved	Accessed	Served
Backscatter	27/51/73/95	20/34/47/60	14/28/36/46	Feb 2003
Passive	298/275/185/233	220/211/150/179	184/173/127/157	Feb 2004
Topology	114/155/163/129	91/129/113/83	70/76/73/51	Jul 2004
Witty	16/16/16/27	12/12/13/17	10/10/11/14	Mar 2008
Telescope	25/29/34/37	16/22/23/21	13/18/19/17	Jul 2009
DNS-RTT	7/10/7/7	6/8/5/2	5/6/4/2	Aug 2006
DDoS	85/92/108/NA	57/62/75/NA	52/53/67/NA	Mar 2010
Total	**572/628/586/528**	**422/478/426/362**	**348/364/337/287**	

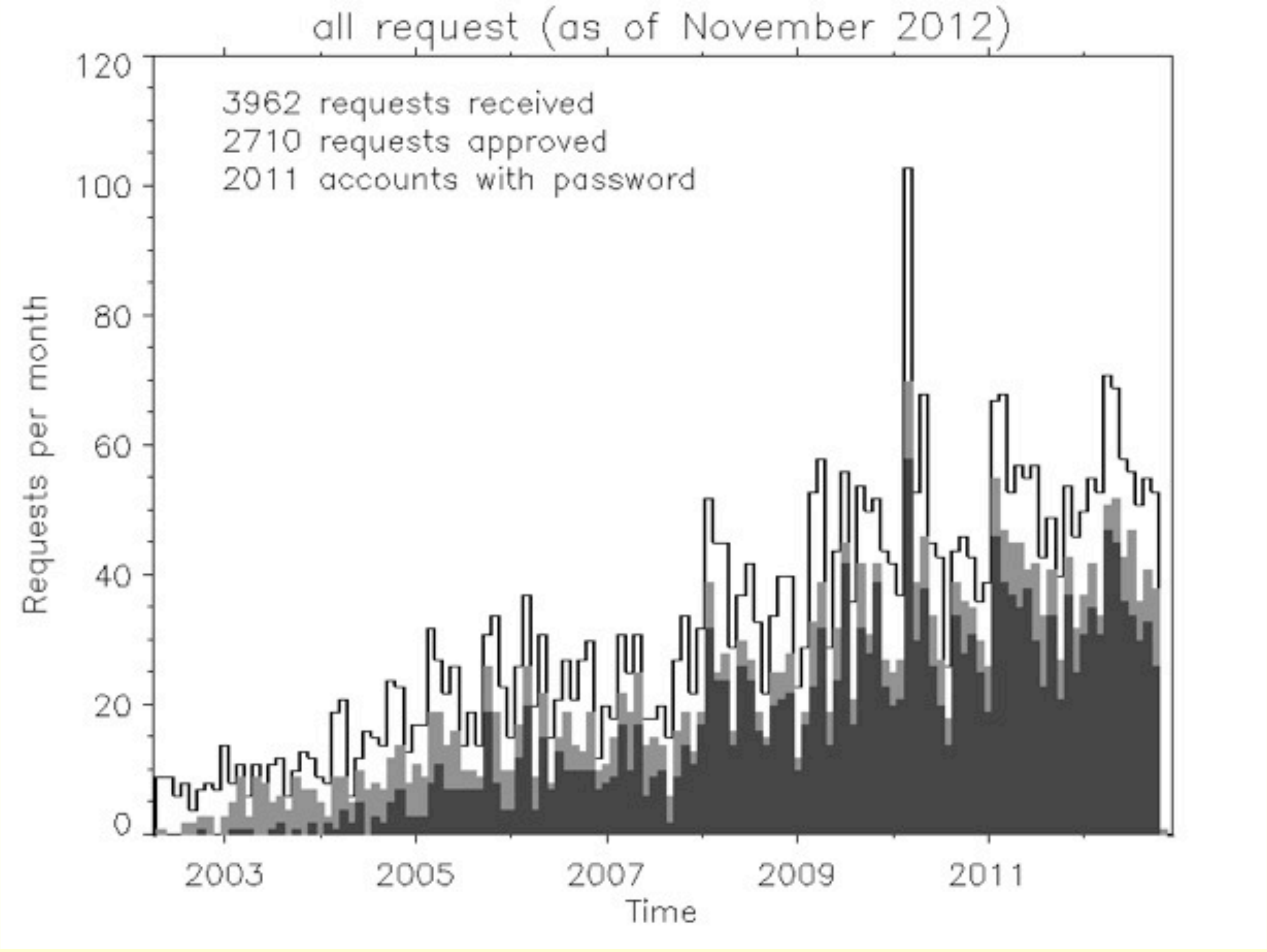
Data request stats

- all requests (cumulative)



Data request stats (cont)

- All requests (monthly)
  - spike (40 requests) in first month of DDoS dataset



Data statistics - online

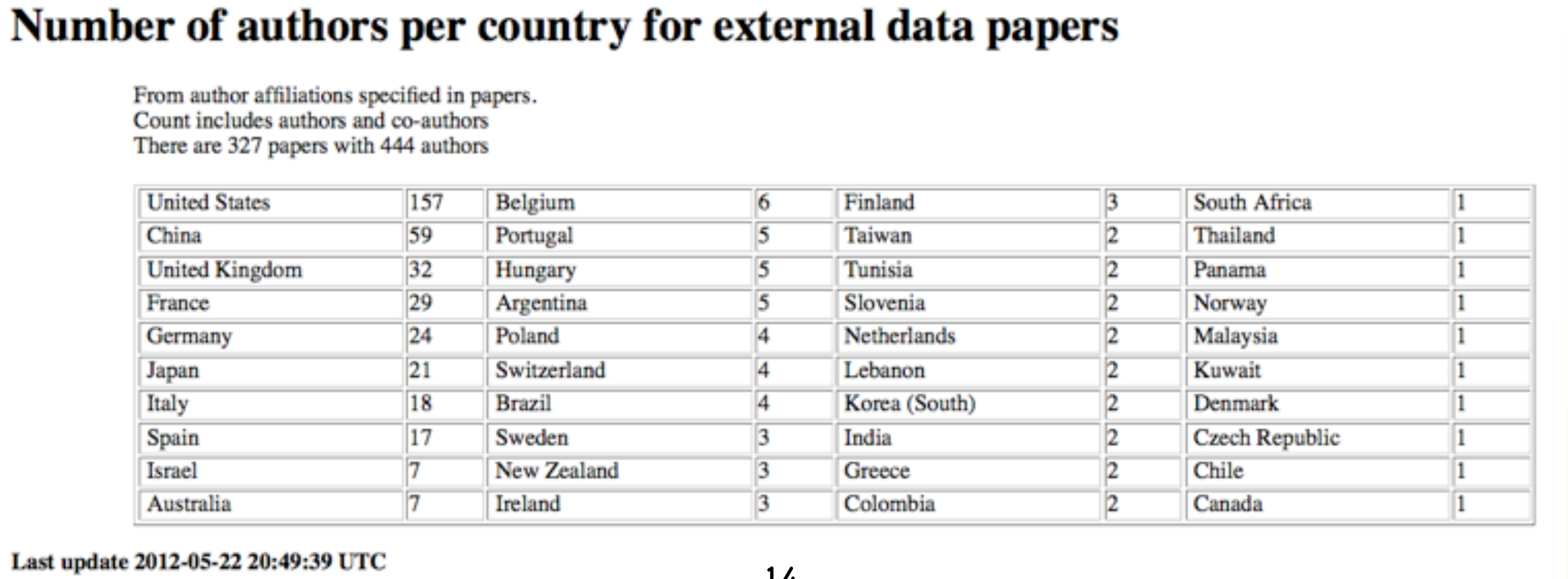
- Report Generator
  - IP-packet-header (traffic) based
  - flows, packet, byte volumes
  - traffic by protocol, port, AS, country, etc
  - http://www.caida.org/data/realtime/passive/?monitor=equinix-sanjose-dirA
- Topology
  - Ark statistics: http://www.caida.org/projects/ark/statistics/index.xml
  - path dispersion (AS and IP), path length distribution, RTT distribution, RTT vs. distance, median RTT per country, ...
- Meta-data for IP packet header data
  - Date, start time, stop time
  - Numbers of IPv4, IPv6, unknown packets
  - Transmission rate in pkts/s, bits/s
  - Link utilization (%)
  - Average packet size
  - Graph of packet size distribution (IPv4 and IPv6)
  - http://www.caida.org/data/passive/trace_stats/

Phase II Data Sets

- UCSD telescope: near Real-Time Telescope Dataset (RTTD)
- topology: Ark data (ongoing)
  - IPv4 Routed /24 Topology dataset
  - IPv4 Routed /24 DNS Names dataset
  - IPv6 Routed Topology dataset
- topology: ITDKs
  - 2012-07
  - 2011-10
  - historical releases back to 2002
- High-speed backbone link: 2007-2012

non-CAIDA publications using PREDICT-related CAIDA data (last search May12)

total	194
backscatter	21
passive-oc48	56
passive-2007	9
witty	14
itdk	13
skitter	57
ark	24 (now in predict)



Number of authors per country for external data papers

From author affiliations specified in papers.
Count includes authors and co-authors
There are 327 papers with 444 authors

United States	157	Belgium	6	Finland	3	South Africa	1
China	59	Portugal	5	Taiwan	2	Thailand	1
United Kingdom	32	Hungary	5	Tunisia	2	Panama	1
France	29	Argentina	5	Slovenia	2	Norway	1
Germany	24	Poland	4	Netherlands	2	Malaysia	1
Japan	21	Switzerland	4	Lebanon	2	Kuwait	1
Italy	18	Brazil	4	Korea (South)	2	Denmark	1
Spain	17	Sweden	3	India	2	Czech Republic	1
Israel	7	New Zealand	3	Greece	2	Chile	1
Australia	7	Ireland	3	Colombia	2	Canada	1

Last update 2012-05-22 20:49:39 UTC

Recent publications

- A. Dhamdhere, M. Luckie, B. Huffaker, kc claffy, A. Elmokashfi, E. Aben, *“Measuring the Deployment of IPv6: Topology, Routing and Performance”,* International Measurement Conference (IMC) November 2012.
- A. Dainotti, A. King, K. Claffy, F. Papale, and A. Pescapè, *”Analysis of a "/0" Stealth Scan from a Botnet”,* submitted to ACM SIGCOMM CCR. (Used Corsaro 8-tuple)
- A. Dainotti, A. King, and K. Claffy, *”Analysis of Internet-wide Probing using Darknets”,* Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), Oct 2012.

Recent publications

- T. Zseby and k. claffy, *”DUST 2012 Workshop Report”,* ACM SIGCOMM Computer Communication Review (CCR), vol. 42, no. 5, pp. 49--53, Oct 2012.
- N. Brownlee, “*One-way Traffic Monitoring with iatmon”,* Passive and Active Network Measurement Workshop (PAM), Vienna, Austria, Mar 2012, PAM 2012.
- k. claffy, “*The 4th Workshop on Active Internet Measurements (AIMS-4) Report*'', ACM SIGCOMM Computer Communication Review (CCR), vol. 42, no. 3, pp. 34--38, Jul 2012.
- k. claffy, *“Border Gateway Protocol (BGP) and Traceroute Data Workshop Report*'', ACM SIGCOMM Computer Communication Review (CCR), vol. 42, no. 3, pp. 28--31, Jul 2012.

Recent presentations

- A. Dainotti, “SipScan: the world scanning itself”, at the *1st International Workshop on Darkspace and UnSolicited Traffic Analysis (DUST 2012)* in May 2012 http://www.caida.org/publications/presentations/2012/dust_sipscan/
- E. Kenneally, *“Illuminating the way for Trusted Darkspace Data Sharing”*, at the 1st International Workshop on Darkspace and UnSolicited Traffic Analysis (DUST 2012) in May 2012 http://www.caida.org/publications/presentations/2012/dust_darkspace_data_sharing/
- A. King, *“Corsaro”*, at the *1st International Workshop on Darkspace and UnSolicited Traffic Analysis (DUST 2012)* in May 2012 http://www.caida.org/publications/presentations/2012/dust_corsaro/
- T. Zseby, *“Comparable Metrics for IP Darkspace Analysis”*, at the *1st International Workshop on Darkspace and UnSolicited Traffic Analysis (DUST 2012)* in May 2012 http://www.caida.org/publications/presentations/2012/dust_metrics_darkspace_analysis/

Recent presentations

- A. Dainotti, *“Extracting Benefit from Harm: Using Malware Pollution to Analyze the Impact of Political and Geophysical Events on the Internet”*, at the ACM SIGCOMM conference in August 2012 http://www.caida.org/publications/presentations/2012/extract_benefit_from_harm_sigcomm_2012/
- A. Dainotti, *“Analysis of Internet-wide Probing using Darknets”*, at the BADGERS 2012 conference in October 2012. http://www.caida.org/publications/presentations/2012/analysis_darknets_badgers/
- M. Luckie, *“CAIDA's AS-rank: measuring the influence of ASes on Internet Routing”*, at the North American Network Operator's Group (NANOG) meeting in October 2012 http://www.caida.org/publications/presentations/2012/caida_asrank_nanog/

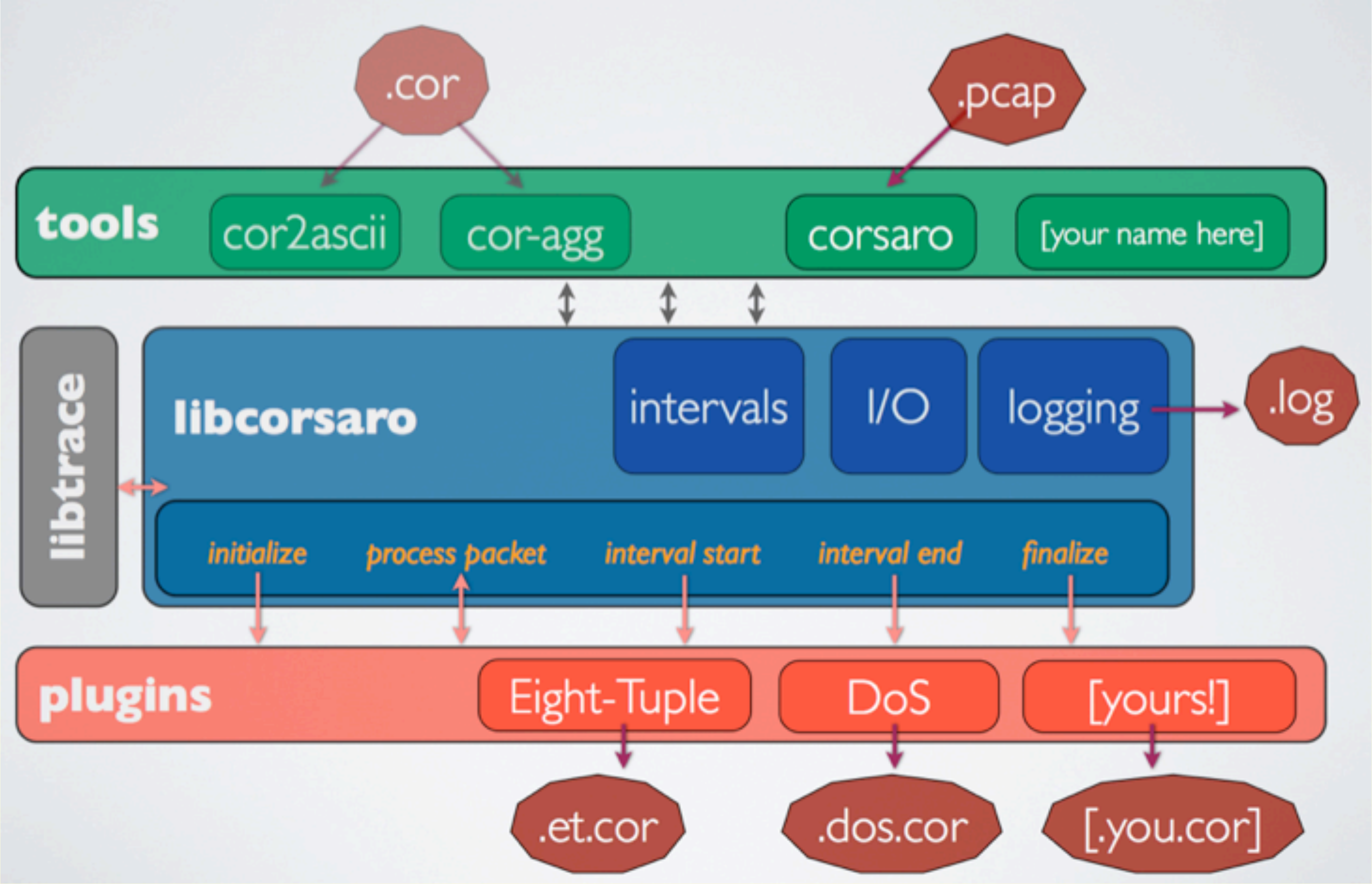
Workshops

- 5th CAIDA-WIDE-CASFI Joint Measurement Workshop
  - 1-2 August 2012 at SDSC, UC San Diego
  - supports a three-way collaboration between researchers from CAIDA (USA), WIDE (Japan) and CASFI (South Korea).
  - 14 presenters
  - presentations at http://www.caida.org/workshops/wide-casfi/1208/
- ISC/CAIDA Data Collaboration Workshop
  - 22 October 2012 in Baltimore, MD co-located with the MAAWG 26th general meeting.
  - showcases novel case studies of network and security data analysis and data sharing
  - data synthesis techniques and technologies
  - 13 presenters

Corsaro - Analysis and Indexing Framework (Update)

- Corsaro documented and available for download http://www.caida.org/tools/measurement/corsaro/
- tools and extensible framework for high-speed packet train ad-hoc analysis, plugins, post-processing data management
- aggregates data into intervals (1-min bins)
- Core plugins
  - Raw pcap
  - 8-tuple flow record balances storage resources and research utility
  - RS DoS - uses heuristics described by Moore et al. in [3] to detect backscatter packets
  - Smee - packet classification next release

Corsaro - Analysis and Indexing Framework


.cor
.pcap
tools
cor2ascii
cor-agg
corsaro
[your name here]
libtrace
libcorsaro
intervals
I/O
logging
.log
initialize
process packet
interval start
interval end
finalize
plugins
Eight-Tuple
DoS
[yours!]
.et.cor
.dos.cor
[.you.cor]