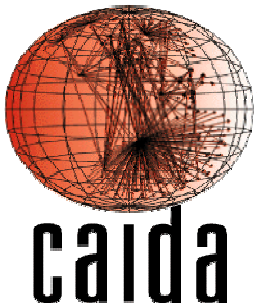
Workload Data Collection Efforts

Colleen Shannon (CAIDA)

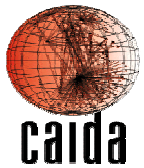
cshannon @ caida.org

www.caida.org



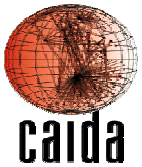
Who collects workload data?

- NLANR
- Sprint ATL
- Abilene/Internet2
- LLBL Internet Traffic Archive
- Netcraft
- Team Cymru
- DShield
- CAIDA
- Wireless data (UCSD and Dartmouth)



Who collects workload data?

- NLANR: National Laboratory for Applied Network Research
  - 90 second packet traces from ~11 locations
  - Collected ~8 times a day
  - IP addresses encoded, freely available
  - A number of large packet traces
    - Recent traces mostly from University of Auckland commodity access link
  - Mostly academic/research traffic
  - http://pma.nlanr.net/PMA/



Who collects workload data?

- Sprint ATL
  - Large traces from Sprint's backbone
    - OC48 traces
    - OC192 traces (measurement cards being developed)
  - Not publicly available; you must work at Sprint to use them
  - Some work on allowing external folks to run scripts to collect data at Sprint, but that's not happening yet…

Who collects workload data?

- Abilene/Internet2
  - MRTG graphs for networks
  - Semi-anonymized (first 11 bits ok, rest scrambled) sampled netflow data from all links publicly available
  - Data collected at 14 locations
  - http://www.abilene.iu.edu/noc.html
  - Academic/research traffic – not commodity Internet

Who collects workload data?

- LLBL Internet Traffic Archive
  - Location for public data submissions
  - http://ita.ee.lbl.gov/html/traces.html
  - Last modified in 2000…

Who collects workload data?

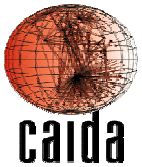
- Netcraft
  - Graphs of:
    - Spread of Internet worms
    - Port scanning
    - High-profile DoS attacks
  - Data not publicly available (as far as I know)
  - http://news.netcraft.com/

Who collects workload data?

- Team Cymru
  - Weekly reports:
    - Compromised machines
    - Open proxies, Spam sources
    - Worm infected machines (Blaster, Slammer, Nachi…)
    - Hijacked DNS
    - Hijacked ASNs

Who collects workload data?

- DShield – distributed internet intrusion detection
  - Aggregation of volunteers reporting data
  - Similar to distributed blackhole/network telescope
  - Parses firewall-type logs into summaries to identify trends and develop better firewall or other blocking activities
  - http://www.dshield.org/intro.php



Who collects workload data?

- CAIDA
  - Passive traces:
    - OC48 and University links
    - OC48 traces available with encoded IP addresses
    - Unencoded OC48 traces available if you visit CAIDA
  - Network Telescope
    - Summarized flow data with encoded IP addresses available in near realtime soon
    - Packet headers/~40 bytes of payload of incoming anomalous traffic (backscatter, worms, scanning, junk); trying to figure out how to make this data available responsibly

Who collects workload data?

- Wireless Data
  - Dartmouth:
    - Several publicly available traces from campus and research wireless networks
    - http://cmc.cs.dartmouth.edu/data/
  - UCSD:
    - Publicly available traces from Sigcomm 2001
    - Currently collecting new data that will probably be publicly available
    - http://ramp.ucsd.edu/pawn/sigcomm-trace/

Who collects workload data?

- Internet Measurement Data Catalog
  - A work in progress! Hoping to have a minimally functional prototype available in the next year…
  - Index of available data
  - Archive of information *about* data, but not the data itself
  - Extensive annotation functionality
  - Ability to link data with published papers
    - Very important for repeatability/verification of research results

Who collects workload data?

- What I wish I could collect…
- What I wish I could share…
  - Everything, but only with folks who won't abuse it
- Stuff I worry about
  - Walking the line between security and utility
  - Keeping track of data that is collected