Archipelago

Measurement Infrastructure

Status and Experiences

Young Hyun

CAIDA
10th CAIDA-WIDE Workshop
Aug 15, 2008

Introduction

* Archipelago (Ark) is CAIDA's next-generation active measurement infrastructure
  * evolution of the skitter infrastructure
* in production since Sep 12, 2007

Outline

* Monitor Deployment
* Four Datasets
* Alias Resolution
* Lessons Learned
* Future Work

Monitor Deployment

* 27 monitors in 21 countries
  * 24 actively probing in 17 countries
  * 3 inactive
    - hardware problem; ICMP rate limiting; IPv6 only
    - Germany, Hungary, Luxembourg
  * 8 in US

Monitor Deployment

* down the road: 38 monitors in 26 countries
  * 4 monitors coming soon (< month?):
    - Canada, US (3)
  * 7 monitors in next 1-6 months:
    - China, Argentina, Italy, South Africa, Pakistan, US (2)

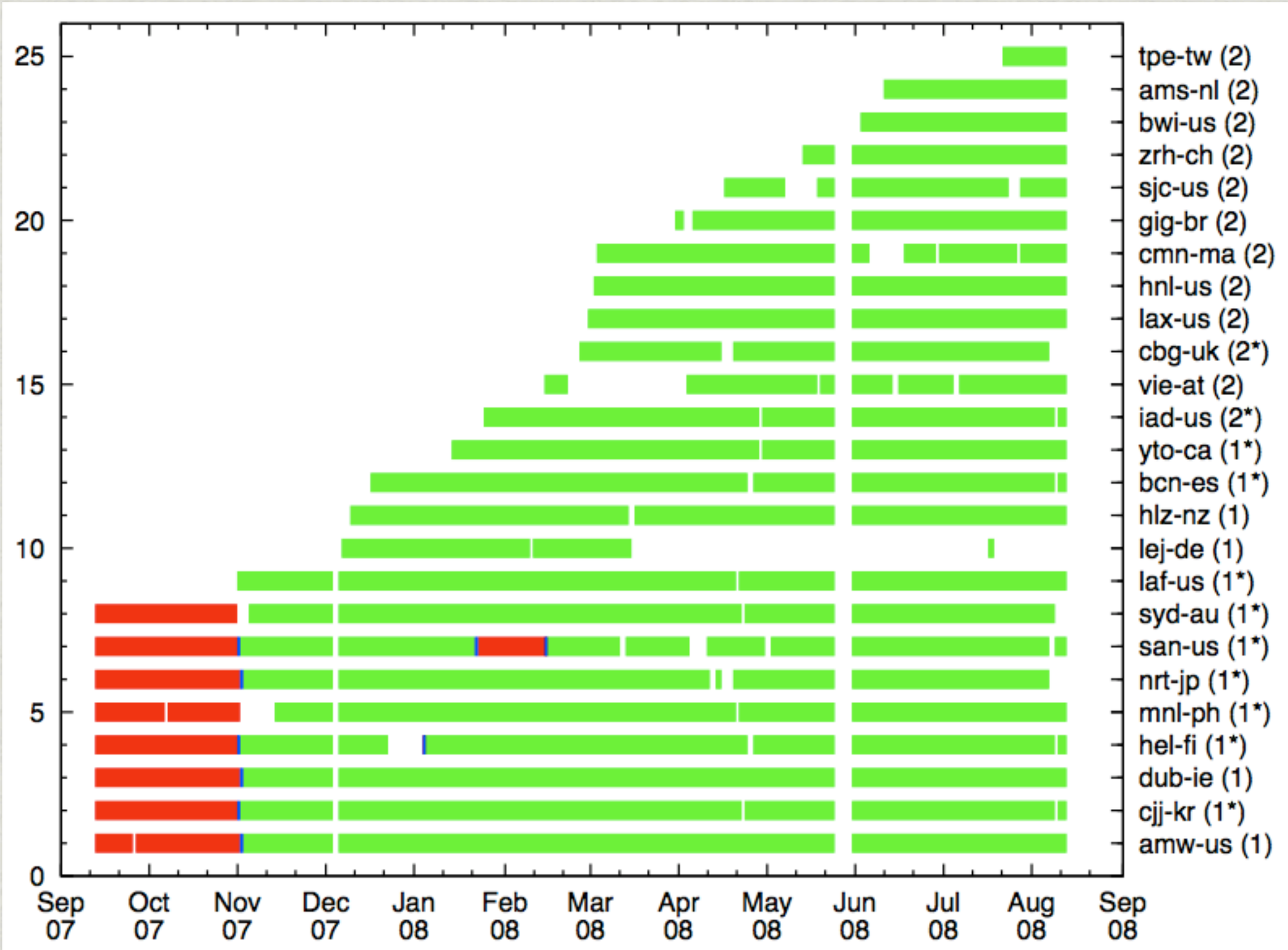
Datasets

* IPv4 Routed /24 Topology
* IPv4 Routed /24 AS Links
* DNS Names
* DNS Query/Response Traffic

IPv4 Routed /24 Topology

- ongoing large-scale topology measurements
  - ICMP Paris traceroute to every routed /24 (7+ million)
  - running scamper
    - written by Matthew Luckie of WAND with funding from WIDE among others
- group monitors into teams and dynamically divide up the measurement work among team members
  - 13-member team probes every /24 in 48-56 hrs at 100pps
  - only one monitor probes each /24 per cycle
- so far, from Sep 12, 2007 to Aug 12, 2008:
  - 1.3 billion traceroutes; 519GB of warts files

IPv4 Routed /24 Topology



red = UDP; green = ICMP Paris; blue = mix

IPv4 Routed /24 AS Links

- AS links from Routed /24 Topology traces
  - map IP addresses to ASes with RouteViews BGP table
  - one AS links file per cycle per team
- statistics:
  - ~29k AS links per cycle per team at present
  - AS links observed over most recent 1, 2, and 7 months (# cycles in parentheses):

	July	June-July	Jan-July
team 1	48.1k (12)	58.1k (26)	89.7k (94)
team 2	47.5k (11)	55.1k (20)	75.4k (48)
combined	56.6k (23)	67.1k (46)	100.1k (142)

These are **lower bounds**: no AS links files generated for some cycles due to bug.

DNS Names

- automated ongoing DNS lookup of IP addresses seen in the Routed /24 Topology traces
  - all intermediate addresses and *responding* destinations
  - using our in-house bulk DNS lookup service (HostDB)
    - can look up millions of addresses per day
- 85M hostnames since March 2008

DNS Traffic

- tcpdump capture of DNS query/response traffic
  - only for lookups of Routed /24 Topology addresses
  - continuous collection of 3-5M packets per day
  - can download most recent 30 days of pcap files
- a broad sampling of the nameservers on the Internet due to the broad coverage of the routed space in traces
- how many nameservers have IPv6 glue records? DNSSEC records? support EDNS? typical TTLs?

Alias Resolution

- Goal: collapse interfaces observed in traceroute paths into routers
  - toward a router-level map of the Internet
- alias resolution work led by Ken Keys

Alias Resolution

- how much topology data should we examine?
  - about time period (window), not quantity
    - last month, 3 months, or year of traces?
  - window must be large enough
    - include topology traversed infrequently or irregularly
    - in Routed /24 Topology dataset, only **one** monitor probes each /24 per cycle
  - window should not be too large
    - may include topology that no longer exists
    - will increase amount and difficulty of processing

Alias Resolution

- how much topology data should we examine?
  - answer depends on the frequency of appearance of addresses and links
  - if an address or link is absent, is it really gone from the topology or did we just fail to sample it?
- based on our analysis, 20-24 cycles seem to be a reasonable window (at 2-3 days per cycle)
  - we examined all team 1 data available on 2008-06-24
    - 108 cycles in 285 days
    - 3.3M intermediate addresses (that is, non-destinations)
    - 5.9M IP links

Alias Resolution

- all techniques have strengths and weaknesses, so we combine them to get the best results
- our plan:
  - run **iffinder** on Routed /24 data
  - run **APAR** on Routed /24 data and iffinder results
  - run **Ally** on final set of aliases, as validation

Alias Resolution: **iffinder**

- written by Ken Keys
- “Mercator technique” described by
  - J.-J. Pansiot and D. Grad. "On routes and multicast trees in the Internet."
  - R. Govindam and H. Tangmunarunkit. "Heuristics for Internet Map Discovery."
- procedure:
  - send UDP packet to unused port on all router interfaces
  - ICMP Port Unreachable response from a **different** address implies that the target address and reply address may be aliases

Alias Resolution: **iffinder**

- we ran iffinder on 23 Ark nodes using 24 cycles of Routed /24 data (team 1 only)
  - each node probed the same 2.3M addresses in random order
  - took about 11 hours per node
- results were similar across all nodes:
  - 1.25M to 1.36M port unreachables
    - 54% to 59% of probed addresses
  - 45k to 50k port unreachables from **different** address
    - about 2% of probed addresses

Alias Resolution: **iffinder**

- combined results:
  - 118k interfaces in 46k interface sets (possible routers)
    - 5% of 2.3M addresses probed were assigned to a set
  - averaging 2.5 interfaces per set
  - 35k sets have 2 addresses
  - 455 sets have 10 or more addresses
  - largest set has 100 addresses

Alias Resolution: **APAR**

- **A**nalytical and **P**robe-based **A**lias **R**esolution
  - M. Gunes and K. Sarac. "Resolving IP Aliases in Building Tracroute-Based Internet Maps."
- procedure:
  - ping router addresses to collect TTLs
    - we can simply examine TTLs in our collected traces
    - use TTLs to rule out pairs of interfaces as aliases
  - identify subnets that router addresses belong to
    - look for common prefixes that do not introduce any contradictions (e.g., loops) in the graph
  - compare paths that cross the same subnet in opposite directions to infer aliases

Alias Resolution: **APAR**

* compare paths that cross the same subnet in opposite directions to infer aliases:

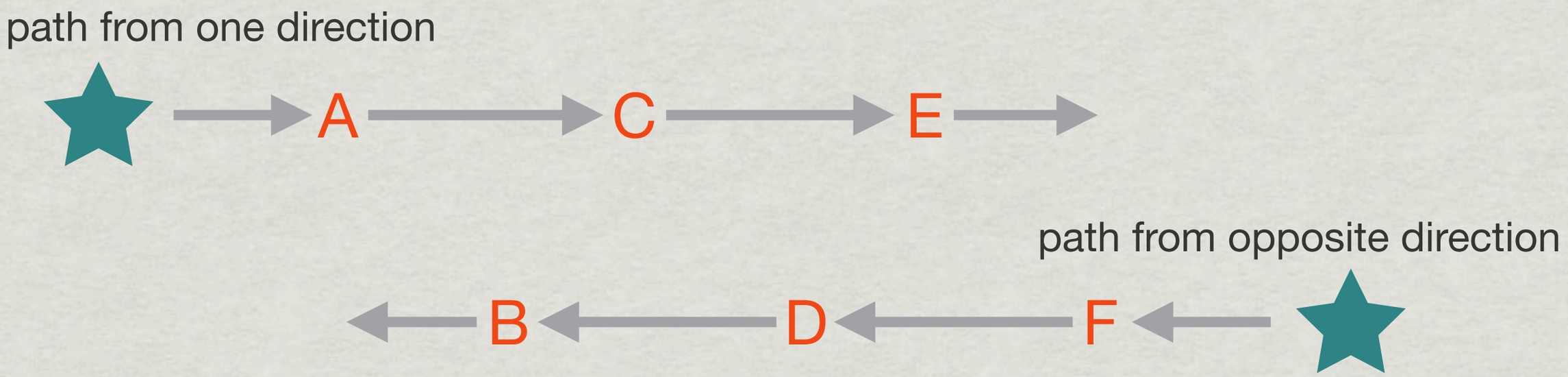
Alias Resolution: **APAR**

* compare paths that cross the same subnet in opposite directions to infer aliases:

path from one direction

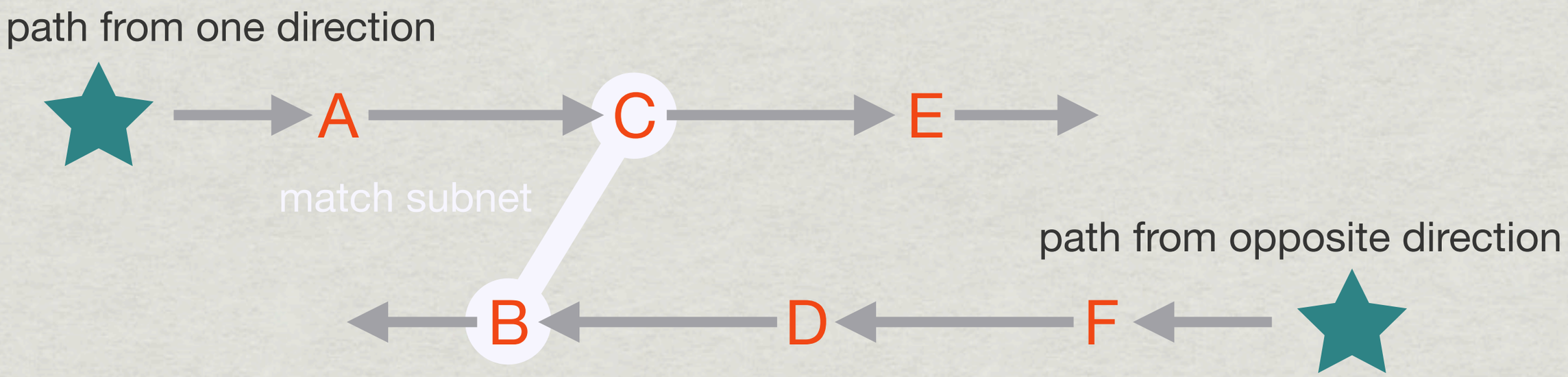
Alias Resolution: **APAR**

* compare paths that cross the same subnet in opposite directions to infer aliases:



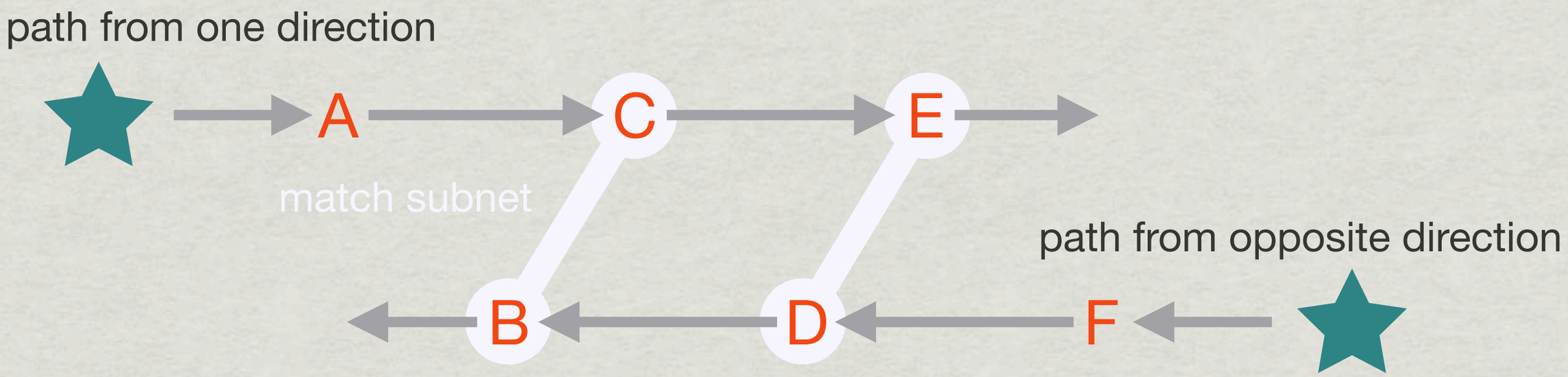
Alias Resolution: **APAR**

* compare paths that cross the same subnet in opposite directions to infer aliases:



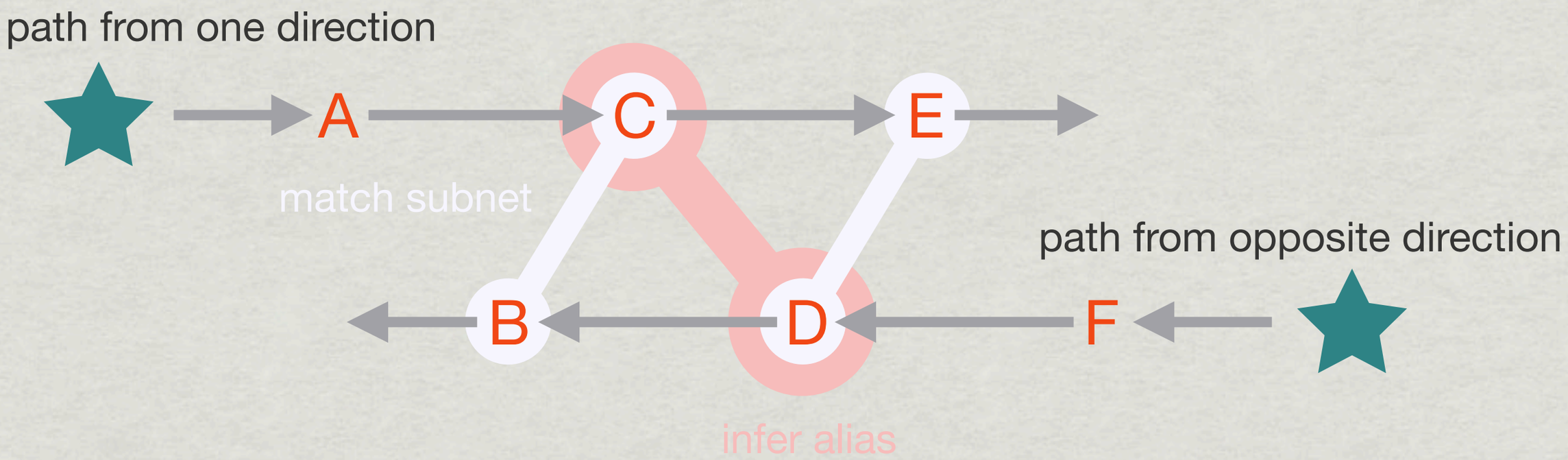
Alias Resolution: **APAR**

* compare paths that cross the same subnet in opposite directions to infer aliases:



Alias Resolution: **APAR**

* compare paths that cross the same subnet in opposite directions to infer aliases:



Alias Resolution: **APAR**

* Ken Keys is writing a scalable implementation that can handle ~100M traces seen in Routed /24 dataset
  * have a way to avoid loading full paths into memory

Alias Resolution: **Ally**

- N. Spring, R. Mahajan, and D. Wetherall. "Measuring ISP topologies with Rocketfuel."
- procedure:
  - for each pair (A,B) of possible aliases, send sequence of probes to A, B, A, B
  - if IP ID of responses are sequential and close, A and B are most likely aliases
  - if A and B each respond with sequential IP IDs, but they do not agree with each other, then A and B are most likely not aliases
- we have too many addresses to probe all N^2 pairs
  - useful to validate alias pairs found by other means

Alias Resolution: **Weaknesses**

- iffinder
  - not many interfaces reply with a different address
  - slight chance for false positives with middleboxes
- APAR
  - needs paths in both directions across several adjacent interfaces
  - relies on traceroute *paths*, which can be incorrect
- Ally
  - without some means of choosing likely pairs (e.g., clustering interfaces by TTL), need to probe $O(N^2)$ pairs
  - some routers do not use sequential IP IDs

Alias Resolution: **Validation**

- acquire list of true router interfaces from network providers and compare
  - beg and plead for information (*might* work)
  - WIDE and IIJ router lists?
- validate APAR and Ally with Internet2 data
  - but not iffinder, because Internet2 routers do not respond to iffinder probes

Lessons Learned

- it takes a **long** time to deploy a monitor
  - tireless work of Emile Aben and Dan Andersen
  - for example:
    - NLANR's AMP infrastructure decommissioned in July 2006
      - worked for >2 years to repurpose boxes for Ark
    - skitter infrastructure decommissioned in Feb 2008
      - still not fully migrated to Ark after 6 months
  - need to
    - (re-)establish contact
      - former contact long gone; spam filters; language hurdles
    - get approval for Ark AUP (Memorandum of Understanding)
    - remote upgrade operating system or ship out new box
    - troubleshoot and fix network connectivity
      - firewalls; broken routing; mysterious rate limiting

Lessons Learned

* expect fAiLUre
  * hardware failures (of course)
    - power supply death; hard drive death
  * software bugs
    - plenty of my own
    - but also helped tracked down
      - FreeBSD kernel crash when passing file descriptors via Unix sockets
      - Ruby memory leak and threading deadlock
  * time jumping backwards under Xen
  * network problems
  * power loss, power loss, power loss

failures and issues since Sep 12, 2007

- 2007-09-12 to 2007-09-18: Ark software issues
- 2007-09-25: amw-us rebooted due to faulty breaker
- 2007-10-06: mnl-ph crashed? (/ not properly unmounted, no sign of controlled reboot)
- 2007-10-15: team-sorter coredump
- 2007-10-30: syd-au unreachable
- 2007-10-31: mnl-ph produce less traces per hour -- rate limiting?
- 2007-11-07: a few hours idling due to issue with switching over to new BGP prefixes
- 2007-11-20: downloading from mnl-ph stopped, though box itself continued probing
- 2007-11-26: cjj-kr stopped providing data
- 2007-12-03: global tuple space wedged??
- 2007-12-22: hel-fi: scamper coredumped
- 2008-01-04: multiple cycles in daily files
- 2008-01-22: san-us accidentally switched to probing with UDP (fixed 2008-02-15)
- 2008-02-08: lej-de crashed
- 2008-02-15: san-us crashed
- 2008-02-15: vie-at temporarily hung due to time going backwards
- 2008-02-16: ark-collector failed for vie-at
- 2008-02-23: lej-de crashed
- 2008-03-06: rebooting san-us to get IPv6 configured properly
- 2008-03-07: scamper-20070523-p8 crashed on syd-au
- 2008-03-08: ark-collector failed for mnl-ph: TypeError exception
- 2008-03-09: ark-collector had transient (that is, recovered) problems for a while for bcn-es
- 2008-03-10: lej-de crashed
- 2008-03-10: san-us crashed -- hard drive seems to be dying
- 2008-03-11: global tuple space wedged??
- 2008-03-14: hlz-nz rebooted due to power outage
- 2008-03-14: lej-de crashed
- 2008-03-15: lej-de crashed again and again and again
- 2008-03-19: ark-collector failed for cmn-ma: TypeError exception
- 2008-04-01: scamper-l produced corrupted gig-br traces
- 2008-04-01: scamper-l crashed on gig-br
- 2008-04-15: updated & restarted nrt-jp & cbg-uk after recompiling 4.6 kernel to have bpf
- 2008-04-18: mnl-ph crashed
- 2008-04-21: updated & restarted syd-au & cjj-kr after recompiling 4.6 kernel to have bpf
- 2008-04-23: bcn-es extended network problems but Ark automatically recovered
- 2008-04-24: vie-at unclean reboot after Ubuntu upgrade
- 2008-04-24: updated & restarted bcn-es & hel-fi after recompiling 4.6 kernel to have bpf
- 2008-04-25: vie-at clock problem suspending processes
- 2008-04-26: gig-br: TimedSFTP got TypeError
- 2008-04-30: updated & restarted cmn-ma after recompiling 4.6 kernel to have bpf
- 2008-05-01: dub-ie high loop rates in Apr 2008 caused by MPLS loops in immediate provider
- 2008-05-01: crap, used "-I 2" and not "-L 1" on dub-ie & vie-at
- 2008-05-08: cmn-ma crashed
- 2008-05-08: seem to have had problems with our CENIC link (or its upstream NLR)
- 2008-05-15: 39% packet loss to nr-jp and 32% to mnl-ph for hours
- 2008-05-18: vie-at unplanned reboot for emergency ssh patching
- 2008-05-20 to 21: gap in san-us data collection due to problems with scamper-trunk
- 2008-05-24: global tuple space wedged
- 2008-05-29: hlz-nz crashed/rebooted?
- 2008-06-04: bcn-es unreachable: earlier switch over to different vlan than planned
- 2008-06-05: vie-at unreachable due to routing loop; vie-at power outage
- 2008-06-05: cmn-ma unreachable
- 2008-06-12: vie-at unreachable
- 2008-06-13: dub-ie crashed/rebooted
- 2008-06-19: cmn-ma rebooted
- 2008-06-29: vie-at rebooted?
- 2008-06-30: cmn-ma multiple power loss
- 2008-07-01: cmn-ma multiple power loss
- 2008-07-04: vie-at rebooted?
- 2008-07-18: lej-de down -- replacement power supply died
- 2008-07-21: downloading from zrh-ch failed due to unexpected Errno::ECONNRESET
- 2008-07-21: san went down; a few Ark stuff affected but data collection continuing
- 2008-07-23: sjc-us and its net seem to be globally unreachable
- 2008-07-26: cmn-ma crashed/rebooted
- 2008-07-28: yto-ca downloading failed with EOFError exception
- 2008-08-04: gig-br downloading failed with Net::SSH::Disconnect exception
- 2008-08-04: global tuple space messed up (as before?)
- 2008-08-05: vie-at rebooted
- 2008-08-07: cmn-ma unreachable from everywhere but box still up
- 2008-08-13: hlz-nz crashed/rebooted

powerfAiLEd

Future Work

- Goals of Ark:
  - make it easy to develop and deploy measurements
    - easy to use communication and coordination facilities
      - Marinda tuple space
    - high-level packet generation, capture, and analysis API
      - inspiration from Scriptroute, Metasploit, Scapy, Racket
  - allow semi-trusted 3rd parties to conduct measurements
    - isolation between users and between measurements
    - enforce policies
      - bandwidth usage, destination selection, type of packets

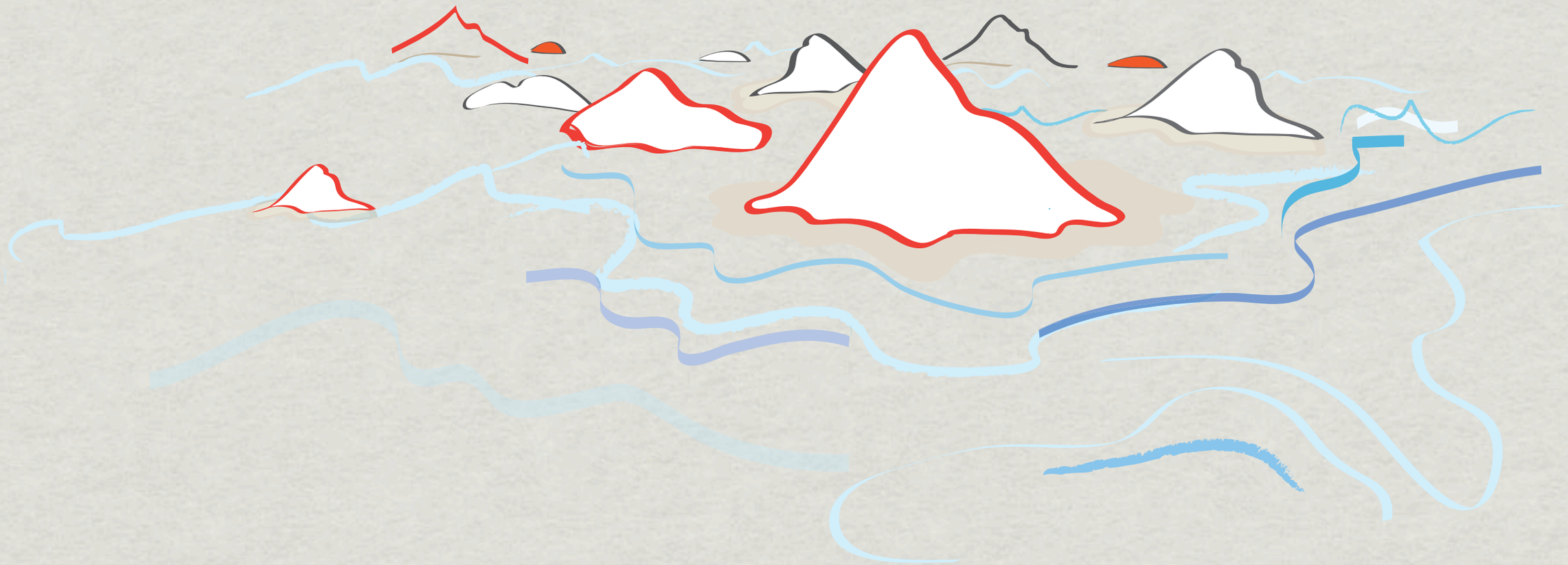
Future Work

- Goals of Ark:
  - make it easy to develop and deploy measurements
    - [x] easy to use communication and coordination facilities
      - Marinda tuple space
    - [ ] high-level packet generation, capture, and analysis API
      - inspiration from Scriptroute, Metasploit, Scapy, Racket
  - allow semi-trusted 3rd parties to conduct measurements
    - [ ] isolation between users and between measurements
    - [ ] enforce policies
      - bandwidth usage, destination selection, type of packets

Future Work

- “large-scale” :-) IPv6 topology measurements
  - 5 deployed monitors currently have IPv6 connectivity
  - more coming
- DNS open resolver surveys?

Thanks!



www.caida.org/projects/ark