PREDICT
Principal Investigator Update

March 2011

Bill Woodcock

Packet Clearing House

PCH Datasets

Phase I:

Routing Topology Dataset (daily snapshots)
End-to-End Quality Dataset

Phase II online now:

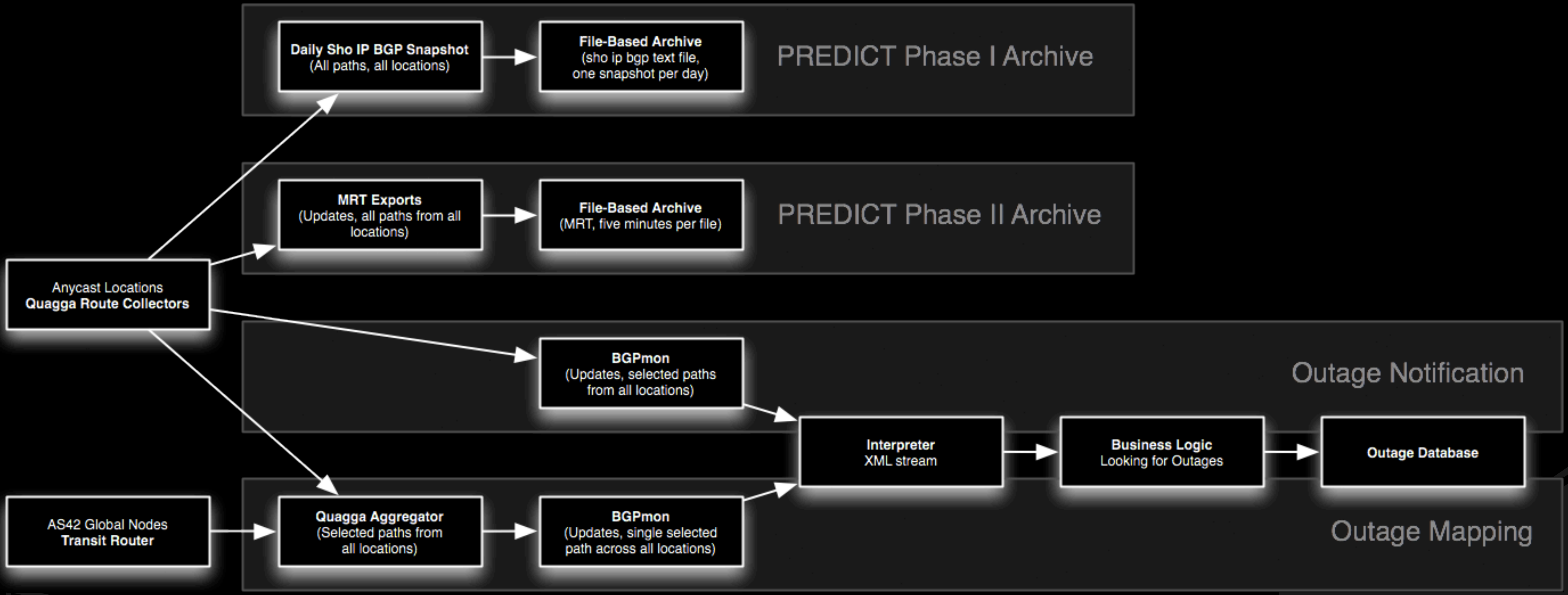
Routing Topology Dataset (real-time MRT updates)

Phase II in process:

Infrastructure topology data
Routing prefix origin inconsistency data
DNS query metadata
Routing outage data

Packet Clearing House

MRT BGP Updates



PREDICT Phase I Archive

Daily Sho IP BGP Snapshot
(All paths, all locations)

File-Based Archive
(sho ip bgp text file, one snapshot per day)

PREDICT Phase II Archive

MRT Exports
(Updates, all paths from all locations)

File-Based Archive
(MRT, five minutes per file)

Anycast Locations
Quagga Route Collectors

Outage Notification

BGPmon
(Updates, selected paths from all locations)

Interpreter
XML stream

Business Logic
Looking for Outages

Outage Database

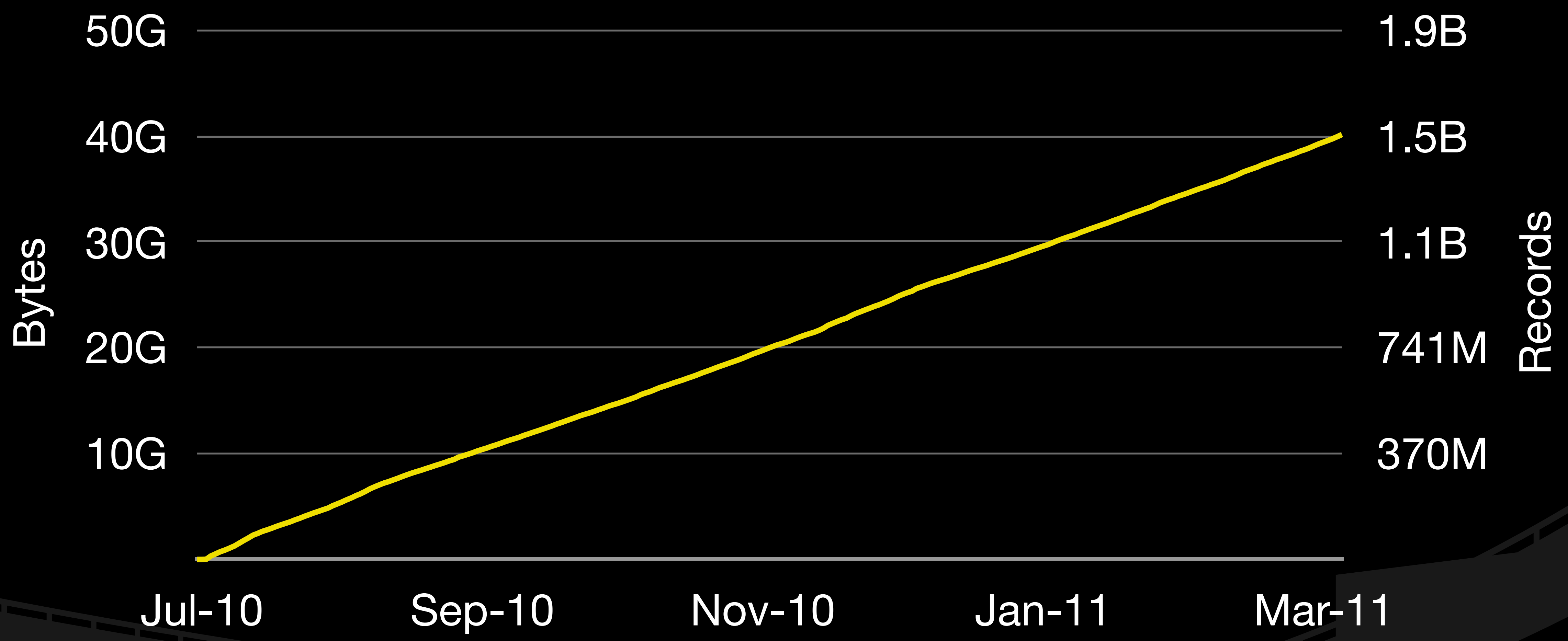
Outage Mapping

AS42 Global Nodes
Transit Router

Quagga Aggregator
(Selected paths from all locations)

BGPmon
(Updates, single selected path across all locations)

MRT BGP Updates



Phase II Datasets

In addition to the MRT-formatted BGP updates, PCH will make four other datasets available through PREDICT in Phase II:

- Internet infrastructure information, including data about Internet exchange points, fiber cable systems, and root and TLD nameservers.
- DNS query metadata (as yet undefined. Suggestions?)
- Routing prefix origin inconsistencies
- Routing outage data.

Packet Clearing House

Infrastructure Data

Infrastructure data is information and metadata about the Internet’s component physical systems. Infrastructure data can be used to analyze the growth properties of the network, interpret observed changes in the network topology, and to correlate real-world organization names, geography, and history with network features as measured and observed from within.

DNS Query Metadata

This is DNS metadata, not an unprocessed dataset. It consists of statistics regarding the quantity and relative frequencies of different types of DNS queries over time, as observed at major authoritative and recursive domain name servers. This metadata can reasonably be used to assess the uptake of protocols like IPv6 and DNSSEC, the spread of code that causes characteristic malformed queries, and general trends in volume and distribution of end-user activity.

Routing Prefix Origin Inconsistencies

This is routing metadata, not an unprocessed dataset. It consists of daily tables of inconsistencies between the expected and observed BGP originating Autonomous Systems for IP prefixes. Inconsistencies are categorized as either major or minor, where major inconsistencies are those in which there is no overlap between the set of expected origin ASNs and the set of observed origin ASNs, and minor inconsistencies are those in which there is incomplete overlap, or the expected and observed ASNs differ but belong to the same organization.

This metadata can reasonably be used to pick out instances of misconfiguration of Internet routers, drift of real-world topology away from baseline over time, hijackings, mergers and acquisitions, “bad actors” in Internet routing security, and other aspects of the Internet routing security environment.

Routing Outage Data

This is routing metadata, not an unprocessed dataset. It consists of logs of Internet outage events, where an Internet outage consists of the BGP withdrawal of the last observed instance of a previously-routed prefix, paired where possible with the next observed reannouncement of that prefix.

This metadata can reasonably be used to quantify the frequency and severity of Internet service outages, and to identify correspondences, correlations, or interdependencies between outages occurring in multiple networks simultaneously or sequentially.

Thanks, and Questions?

Bill Woodcock
Research Director
Packet Clearing House
woody@pch.net