Project Status Report

Reporting period: 04/01/2024 - 09/30/2024
Project title:
Mid-Scale RI-1 Design Project (M1:DP):
Designing a Global Measurement Infrastructure to Improve Internet Security (GMI3S)
OAC-2131987

Principal Investigator: kc Claffy, Bradley Huffaker (UCSD), David Clark (MIT)
Project Manager: Elena Yulaeva
Lead Institution: CAIDA, UCSD
Other Institutions: NSRC (U Oregon), MIT
Cognizant PO:

Table of Contents

1. Summary of project status

A brief summary of the project's overall status on technical progress, cost and performance.

Award Duration	Start date: 10/01/2021	Planned close out: **09/30/2025***
Project Finish Date	Planned Early Finish:	Estimated Early Finish:
Project %-complete	81% (M35)	

2. Near-Term Milestones

Below are milestones with the scheduled dates or actual/forecast/revised (A/F/R) dates that are in current reporting period and until the end of the project, and milestones (with past scheduled dates) that are delayed to future reporting period. (**Completed deliverables have bold font dates**. Red dates have slipped from their original schedule completion dates.) Note that we found the need to revisit some of the milestones as we learned more about other dimensions of the project and as the ecosystem evolved. The date marked with "R" indicates that the WBS element was reopened due to new requirements and was subsequently completed by that day.

WBS	Subsystem	Milestone	Scheduled Date	Actual date (A) /Forecast Date (F)
1.1	**Design Infrastructure for Data Acquisition**			
	1.1.3	Monitors Requirements documented	10/31/2023	10/31/2024 (F)
	1.1.3.7	DNS monitoring needs compiled	04/30/2024	10/31/2024 (F)
	1.1.4	Monitor specification final report	09/30/2024	03/31/2025 (F)
	1.1.4.2	Monitor specification report Draft (2) for community feedback	05/31/2024	**08/31/2024 (A)**
	1.1.4.3	Post report for public comments	08/31/2024	01/31/2025 (F)
	1.1.4.4	Incorporate public comments & publish final document	09/30/2023	03/31/2025 (F)
	1.1.5	Monitor software Prototyped	03/31/2024	12/31/2024 (F)
	1.1.5.5	DNS data monitoring software prototyped	03/31/2024	12/31/2024 (F)
	1.1.7	Evaluation report of Data Acquisition Component published	04/30/2024	07/31/2025 (F)

	1.1.7.3	Evaluation and enhancement of Data Acquisition Component finished	09/30/2024	05/31/2025(F)
	1.1.7.3.1	BGP data acquisition (GILL) evaluated and enhanced	03/31/2024	03/31/2025 (F)
	1.1.7.3.2	Active measurements evaluated and enhanced	09/30/2024	05/31/2025 (F)
	1.1.8	Virtualization capabilities prototyped	03/31/3024	1/31/2025 (F)
	1.1.8.3	Virtualization capabilities documented and evaluated; report published	03/31/3024	11/30/2024 (F)
	1.1.8.5	Scamper containerization completed	NEW	1/31/2025 (F)
	1.1.9	Vulnerabilities and Data Needs Report	02/28/2023	07/31/2025 (F)
	1.1.9.9	Newly identified dataset added (NEW)	NEW	04/30/2025 (F)
	1.1.9.10	Vulnerabilities and Data Needs Report YR4 revision published (NEW)	07/31/2025	07/31/2025 (F)
1.2	**Design Infrastructure for Data Management**			
	1.2.2	Data storage systems specifications published	**3/31/2023**	**08/31/2024 (A)**
	1.2.2.3	Community feedback incorporated into document	**06/30/2023**	**08/31/2024 (A)**
	1.2.3	Data and Metadata Standards Final Report	08/31/2024	07/31/2025 (F)
	1.2.3.1	Develop metadata schema for existing CAIDA datasets	10/31/2023	11/30/2024(F)
	1.2.3.4	Annual revision of data and metadata specifications	ongoing	02/28/2025(F)
	1.2.3.5	Report on the state-of-the-art metadata generation approaches (NEW)	NEW	07/31/2025 (F)
	1.2.4	Specification of tools for data curation and documentation	08/31/2024	07/31/2025 (F)
	1.2.4.3	Specification of tools for data curation and documentation, report, annually updated	ongoing	**03/31/2024(A)** 03/31/2025 (F)
	1.2.4.4	Final report on specification of tools for data curation and documentation	NEW	07/31/2025 (F)
	1.2.5	Design data and metadata API for access to various datasets	09/30/2024	7/31/2025 (F)

	1.2.5.2	Increase the number of supported data sources, including non-CAIDA datasets	03/31/2024	07/31/2025 (F)
	1.2.5.4	Enhance and document data and metadata APIs, update annually	ongoing	07/31/2025 (F)
	1.2.5.4.3	Year 3(&4) Update of data and metadata APIs	09/30/2024	03/31/2025 (F)
	1.2.5.4.4	Final report on data and metadata APIs published	NEW	07/31/2025 (F)
	1.2.6	SDK libraries developed	09/30/2024	07/31/2025 (F)
	1.2.6.3	Libraries evaluated and enhanced, report published	09/30/2024	07/31/2025 (F)
	1.2.7	Tools for additional data sources integration created	10/31/2023	03/31/2025 (F)
	1.2.7.1	AS path annotations implemented	03/31/2024	03/31/2025 (F)
	1.2.7.2	Report on LLM implementation	NEW	03/31/2025 (F)
1.3		**Design Infrastructure for Broad Usability**		
1.3	1.3.1	Data discovery tools prototyped	07/01/2023	06/30/2025 (F)
	1.3.1.5	Report on integration of automated meta-data/data citation creation into catalog.	03/31/2024	06/30/2025(F)
	1.3.2	Software for disclosure control developed	03/31/2024	01/31/2025 (F)
	1.3.2.7	Resource Portal prototype deployed (NEW)	NEW	01/31/2025 (F)
	1.3.3	Report on Policy tools	03/31/2024	03/31/2025 (F)
	1.3.3.8	New agreements designed and shared	10/31/2023	01/31/2025 (F)
	1.3.4	Case studies on Extensibility	03/31/2024	10/31/2024 (F)
	1.3.4.2	State of Internet report created and shared (1.3.4.2 Conduct meetings and compile data to create community-authored "State of the Internet report")	9/30/2023	10/31/2024 (F)
1.4		**Infrastructure for Outreach**		
	1.4.2	Virtual collaboration environment prototyped and evaluated	**09/30/2024**	**09/30/2024(A)**
	1.4.2.2	Virtual collaboration environment evaluated and improved	**09/30/2024**	**09/30/2024(A)**
	1.4.3	STEM workforce task completed	9/30/2024	02/28/2025 (F)

1.4.3.3	Video tutorials on nodes deployment and management created	3/31/2022	02/28/2025(F)
1.4.4	Quarterly calls conducted, minutes shared	ongoing	**04/30/2024(A)** **07/31/2024(A)** 10/31/2024(F) 01/31/2025(F)
1.4.5	Project Presentations	ongoing	**09/30/2024 (A)** 09/30/2025(F)

3. Executive Summary

This progress report introduces the latest advancements in the CAIDA NSF MSRI design project, which aims to develop the next generation of Internet measurement infrastructure to enhance the security and utility of Internet measurements. Our project focuses on creating innovative platforms and tools for data collection, curation and utilization, particularly targeting data related to the security vulnerabilities within the packet carriage layer of the Internet, which often lead to significant harm. This year we have developed or iterated on preliminary specifications for the required infrastructure components and developed methodologies designed to optimize data collection and storage. We have also explored current and potential approaches to data analysis and visualization, addressing the needs for standardization, interoperability, AI readiness, and compliance as we advance in our design and prototyping phase.

In our previous M25-30 biannual report, we summarized our efforts and outlined our design of the infrastructure by component. That report also provided an executive summary of the state-of-the-art and our progress by each data component (traffic data, routing data, active measurement, DNS data), followed by an update on tasks that span all data types. We do not repeat that material here,but include that interim report; the summary starts on p.6. Here we summarize developments in the six months since that report.

Traffic: One-Way: Telescope Data Monitor
We increased the networking and compute capacity, as well as monitoring capacity, of the UCSD-NT. We prototyped a new Kafka-based deployment that could provide access control, encryption, and easy fan-out scaling to share data directly with multiple partners, and started publishing telescope data to the Kafka broker, in order to test and evaluate performance of the design. We designed and experimented with a new dual stack (IPv4 and IPv6) telescope, which led to a new CIRC proposal submitted in September.

Traffic: Two-way: Internet backbone 100Gbs monitor
We provided access to our new Anonymized Two-Way Passive Trace dataset, captured on a 100 GB link between Los Angeles and San Jose for nine beta users. We prioritized users with significant experience with network traffic analysis, who demonstrated a clear plan for how they will use the dataset, and who committed to feedback and participation throughout the evaluation period. We are analyzing feedback from those users and improving our data curation and sharing pipeline. (Data request page.)

Interdomain Routing (BGP) data collection infrastructure

In January 2024 we submitted an ACM SIGCOMM 2024 paper outlining our new BGP monitoring platform design. This paper was awarded Best Paper at ACM SIGCOMM 2024 (https://www.acm.org/conferences/best-paper-awards). The current system is prototype-deployed and documented at https://bgproutes.quest. Peering is entirely automated via a web form, and the GILL prototype has adopted over 100 peers (http://185.216.75.11:5050/). Our collaborators and lead authors on this infrastructure component are now investigating operationalization and sustainability options for the instrumentation, including the option of integrating it into our Implementation Phase.

To support expected future needs of this platform, we designed a new methodology to infer geolocation semantics in BGP communities. We submitted this work to IMC24. We discovered several research tasks still required in this area and are training a postdoc to write a CNS-Core proposal to pursue them.

NSRC's RouteViews team designed and produced a software tool to notify peers if their BGP session goes down, which leverages PeeringDB's contact information. NSRC has completed telemetry of Kafka performance and topics served and is using it in operational dashboards. The RouteViews team has released a beta version of their API (https://api.routeviews.org) aimed at network operators and researchers who require access to up-to-date RouteViews data for monitoring the global routing system.

Active Measurement platform

We made progress on implementing our redesigned Ark node, including: containerization, automated node initialization, modern and automated certificate management, deployment of mechanism to support authenticated operation of nodes behind NATs, new tools for management, upgrading, issue tracking, deployment, and communication with hosting sites We also evaluated the ability of deployed nodes to perform controlled spoofing experiments, which would enable security-related measurements such as ROV (a BGP routing security technology) deployment assessments. We expanded deployment for evaluation of this infrastructure component, to Internet2, DREN, and an international NREN.

We achieved a mutually signed MoU to support deployment of our new IXP-based Ark node, including routing, management, and containerization software to support this new architecture.

We completed the redesign of CAIDA's Internet Topology Data Kit software to support automation and improvements in accuracy of annotations. We integrated these changes into the ITDK 2024 data set.

Domain Name System (DNS) data platforms

We completed an analysis of DNS monitoring requirements (**Milestone 1.1.3.7**), which resulted in a submission to ACM Internet Measurement Conference (DarkDNS: Revisiting the Value of Rapid Zone Update) that was accepted and will be presented the first week of November 2024. We presented the results of this work to the ICANN Board Technical Committee meeting for consideration. We also began a new collaboration with the University of Tennessee to evaluate our prototype DNS TLD zone file indexing infrastructure for use in their DNS abuse research.

Data Acquisition Component Evaluation

A strong focus of this (3rd) year of our project has been to undertake extensive evaluation of our data acquisition components via third-party use of the infrastructure and data sets as described in detail below. In each case the activity led to either a change to the design/workflow of data set generation, e.g., to create smaller datasets in response to researcher quest, or to a submission to a peer reviewed conference as an example of how the infrastructure can serve the needs of security researchers (see the Presentations and Publications section). We conducted our own comparative analysis of topology coverage of our infrastructure relative to what one could achieve using cloud vantage points, finding limitations in coverage in both instances, largely related to geographic diversity of vantage points.

Supporting infrastructure for data management

e responded to the National AI Research Resource (NAIRR) Pilot Demonstration Projects DCL, with a supplement proposal to use newly deployed NAIRR resources at SDSC (SDSC's LLM) to automate the extraction of links between data usage and tools from scientific publications. We had two other proposed ideas related to this project that NAIRR leadership encouraged us to submit to later DCLs.

Our paper, A survey on Packet Filtering, which we performed to inform the design of our traffic capture instrumentation, was accepted for publication in the October 2024 issue of the ACM SIGCOMM Computer Communication Review.

We began the process of redesigning our AS Rank BGP data analysis infrastructure to be able to show exactly which observed AS paths we used at which steps of the inference process. This task is in response to repeated community requests for the details behind our inferences.

Data and Metadata APIs

We updated, presented at the workshop, and announced for beta-evaluation, our FANTAIL system for extracting specified subsets of traceroute data.
We updated our BGP2Go platform to include authentication behind CAIDA's Single Sign-On system, one of the first popular services using our newly designed authorization framework. We completed a comprehensive demonstration of a typical use-case of BGP2GO's API.

Outreach and engagement.

In June we convened the third GMI-AIMS workshop. The workshop goal was to review Year 3 GMI3S progress, discuss the current state of the measurement infrastructures, evaluate the GMI3S design and start developing a plan for the Implementation phase of the project. (**WBS 1.4.1 and 1.5.3**)

We have published four blog posts and nine (including IMC 2024) publications since April 2024. We held weekly, monthly, and quarterly meetings with collaborators and stakeholders, as detailed in the report.

Presentations and publications

By the end of the reporting period we had published 27 papers and made 47 presentations
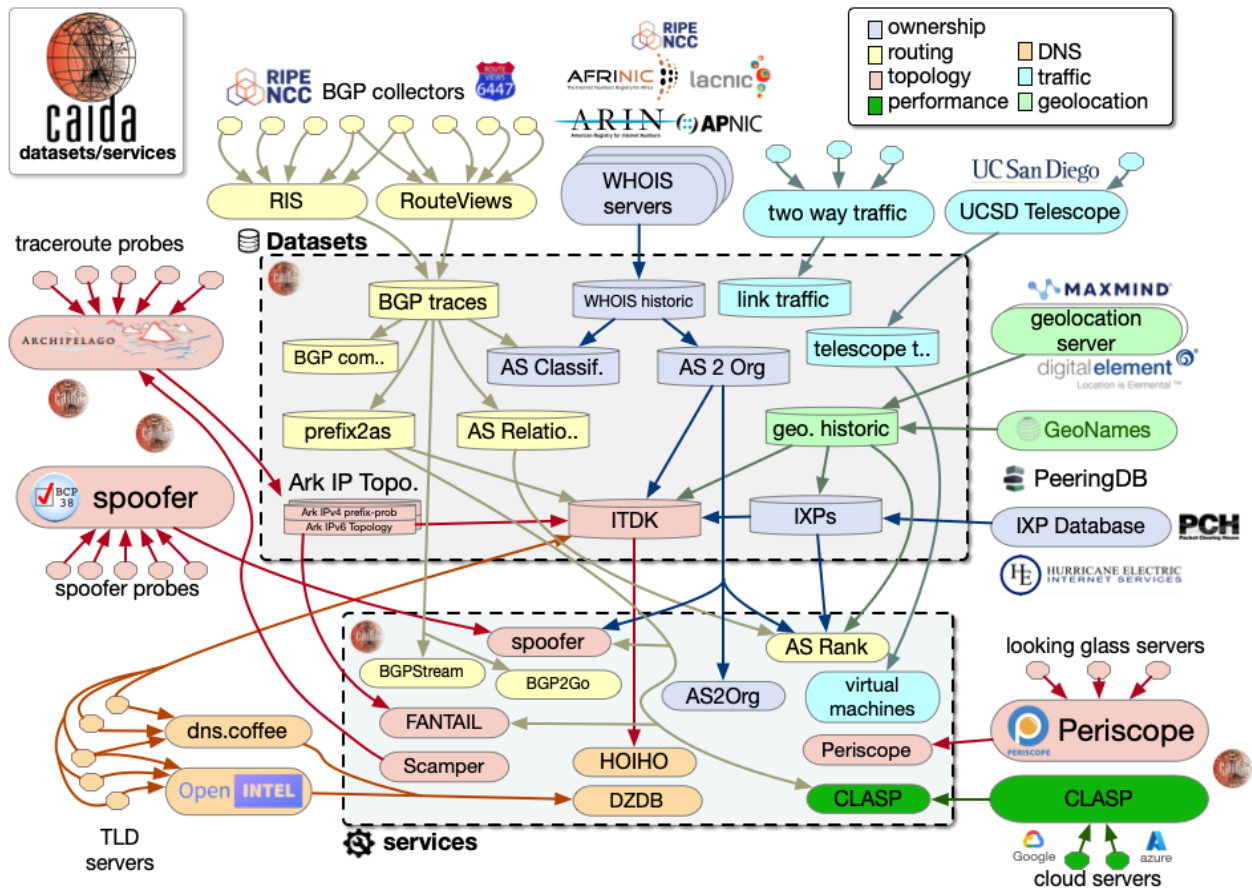https://gmi3s.caida.org/publications/#gmipublications

Figure 1: *GMI Data acquisition components (Task 1)*