Software Engineering Institute

Smart Collection and Storage Method for Network Traffic Data

Angela Horneman
Nathan Dell

September 2014

TECHNICAL REPORT
CMU/SEI-2014-TR-011

CERT® Division

http://www.sei.cmu.edu

Carnegie Mellon University

This material is based upon work funded and supported by the Department of Homeland Security under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center sponsored by the United States Department of Defense.

This report was prepared for the
SEI Administrative Agent
AFLCMC/PZM
20 Schilling Circle, Bldg 1305, 3rd floor
Hanscom AFB, MA 01731-2125

DM-0001506

Table of Contents

List of Figures

List of Tables

Acknowledgments

We would like to thank Soumyo Moitra and Rhiannon Weaver for input into the method and mathematics for the examples. Thank you to everyone in the CERT Division who read the report and provided feedback, especially Sandra Shrum in Technical Communication.

Abstract

Captured network data enables an organization to perform routine tasks such as network situational awareness and incident response to security alerts. The process of capturing, storing, and evaluating network traffic as part of monitoring is an increasingly complex and critical problem. With high-speed networks and ever-increasing network traffic volumes, full-packet traffic capture solutions can require petabytes of storage for a single day. The capacity needed to store full-packet captures for a time frame that permits the needed analysis is unattainable for many organizations. A tiered network storage solution, which stores only the most critical or effective types of traffic in full-packet captures and the rest as summary data, can help organizations mitigate the storage issues while providing the detailed information they need. This report discusses considerations and decisions to be made when designing a tiered network data storage solution. It includes a method, based on a cost-effectiveness model, that can help organizations decide what types of network traffic to store at each storage tier. The report also uses real-world network measurements to show how storage requirements change based on what traffic is stored in which storage tier.

1 Introduction

Network security monitoring (NSM), or capturing and inspecting network traffic for unexpected and malicious activity, is a necessary part of defense-in-depth security solutions. While many security solutions work to prevent security incidents, none are 100 percent effective. NSM helps identify when solutions have failed to identify or prevent incidents, the consequences of those failures, and how to remediate the incidents and improve the prevention mechanisms [Bejtlich 2013]. The methods used for NSM also contribute to network situational awareness: the systematic gathering, analysis, and interpretation of data from local and remote networks, regarding structure, applications, traffic, and resources, to produce actionable information for decision making in network operations and defense.

NSM works by capturing network traffic, either as direct copies of what a sensor sees on the wire (or other transmission media), extracted pieces of the transmitted packets, or summary information on the network flows. Sensors, either hardware- or software-based, capture raw data that is not human readable and must be processed to provide any benefit to analysts. Processing requires that the data be stored in an accessible location secured from unauthorized access. Availability of storage capacity limits how much traffic can be collected and how long it can be kept. While it may be ideal to permanently store a copy of all traffic that ever crosses a network, it is not technologically or financially feasible. Permanent storage would require huge amounts of physical memory and make analysis of the data difficult.

Large networks transmitting over 10 Gbps can generate terabytes of NSM data for each day of collection. Even home networks that implement monitoring may produce many gigabytes of data to store each day. A single Netflix user can generate up to 4.7 GB of network traffic in just one hour [Netflix 2014]. In addition, bandwidth utilization per user is increasing rapidly. As organizations increasingly implement various types of cloud solutions, engage in voice over IP (VoIP) calls, and use video, their network traffic increases as well. The compound annual growth rate for bandwidth utilization for organizations through 2017 is projected to be between 21.7% and 33.7% [Delcroix 2013]. The capacity of commercially available storage media has increased and storage hardware prices have decreased dramatically in the past several years, but the purchase price of these items is only part of the cost of ownership. Managing the hardware's performance, security, and upkeep can cost twice the purchase price each year [Chou 2013]. Buying a $10,000 network-attached system for storing traffic capture data could incur total storage-related costs of $50,000 in just the first two years of the system's life.

With these large amounts of data flowing through high-bandwidth networks, capturing the traffic at wire speed also becomes an issue. High-bandwidth environments may require special network capture cards [Zseby 2009] or custom solutions [Banks 2013]. However, commercial capture cards can cost thousands of dollars [European Information Security Summit 2014], and custom solutions require considerable technical expertise to create, install, and maintain.

Techniques for sampling and filtering traffic allow standard capture solutions to be used even in high-capacity environments [Zseby 2009]. *Sampling* refers to the selection of network packets by capturing only a portion of the packets that traverse a network at random, by choosing one packet

after some number of packets pass the sensor, or by selecting packets for set intervals of time. *Filtering* refers to the selection of network packets by rule matching. Before capture, the filter checks each packet against some specified criteria, and if the packet matches, the filter captures it.

Sampling and filtering both help with another problem. The more data that is stored, the greater the analysis effort and time needed to find the useful information that would identify a security breach or help resolve an incident. Unfortunately, inadequate sampling and poorly designed filtering can be detrimental to analysis. By definition, sampling excludes some part, possibly a large portion, of network packets that might be useful to analysis. In the worst case, sampling could miss all the packets that would identify or describe an incident. Likewise, filtering without a good understanding of the packets that analysts need in order to understand incidents can make the captures useless.

The goal of this report is to provide recommendations and considerations for determining what types of traffic to capture, how much information about the traffic to store as different tiers of data, and for how long. We present what information the different tiers of data can provide, give guidance for determining what parts and types of traffic should be stored to meet policies and service level agreements, and show the benefits of using smart collection techniques over mass collection techniques.

1.1 Network Security Monitoring Background

NSM is a tool that can help with multiple facets of network defense. Traditional uses of NSM are incident response and forensics, but it can also help with post facto anomaly detection and general network situational awareness. In general, the types of data needed for each of these purposes are very similar, whether the organization uses the information primarily in response to alerts (an alert-driven organization) or analysts use the information to actively hunt for anomalies. The major differences in the types of data needed are the levels of detail and the time frames of data that each requires.

All NSM activities are concerned with who or what is communicating over the network, how those communications are taking place, and the purpose (or results of) those communications. Each of the network capture tiers addresses all of these topics, though at varying levels of detail. Analysts for each activity may have different views of the importance of the level of detail necessary to obtain the results they need. Because an organization should conduct each activity—anomaly detection, incident response, forensic investigation, and situational awareness—the organization's capture solution must meet the needs of the activity that requires the most detailed information.

Organizations need to detect and resolve potential issues as soon as possible, so anomaly detection and incident response are ideally near-real-time activities. They need data from the time frame of the last several hours or possibly the last few days. Forensic investigations and situational awareness are more long-term, hindsight activities. To get to the root of an incident or understand what goes on in a network may mean looking at data from weeks, months, or even years ago, as has been the case with some advanced persistent threats [Mandiant 2013]. As with the level of detail, organizations must store data for a time frame that supports the needs of the longest reaching activities: forensic investigations and situational awareness.

The level of detail, storage time frame, and storage capacity all interact in an NSM system. The flow of traffic across an organization's network continuously deposits traffic capture data into a set capacity for storage. Once the storage is at capacity, the influx of new traffic capture data pushes out old data. The balance between the rate of incoming data, the size of the capture data, and the storage capacity determines how long capture data may be retained. Compared to smaller data captures, larger data captures can be stored for shorter periods of time before being pushed out by newer data; the more detailed the capture data, the less network traffic history can be represented, until more storage capacity is added.

Storage time needed for each type and tier of network traffic depends on how the organization wants to control the ongoing costs of packet capture. While the cheapest option may appear to be doing no or very short-term storage of captures, this approach considers only the explicit costs of the energy, personnel effort, and hardware required for capture and storage. Capture solutions both consume budget dollars and reduce losses. Consequently, the actual cost of a system is much more complicated than the price tag of its hardware and maintenance. Examples of the budget-consuming aspects of a solution include

- software licensing (if any—there are many open source options)
- hardware requirements (e.g., taps, memory, storage)
- system maintenance
- analyst training and highly skilled workers
- protection for sensitive data
- possible legal requirements for capturing certain forms of data, such as compliance with electronic discovery regulations

Examples of the loss-reducing aspects of a solution include

- the reduced cost of incident response when captured data exists versus when it does not
- the ability to learn from incidents and close calls to prevent them in the future
- documentation that can aid in attribution and criminal prosecution of incidents
- fulfillment of compliance requirements or industry standards that could otherwise result in fines or other penalties
- decreased liability
- increased understanding of the network with network situational awareness
  - indicating areas that need security improvements
  - allowing for more effective improvements during upgrades
  - detecting network anomalies
  - planning for future architecture considerations

1.2 Tiered Storage Definition and Descriptions

With different types and amounts of information that various monitoring products collect come different benefits and costs. There are three basic tiers of storage: full packet, network flow, and augmented flow.

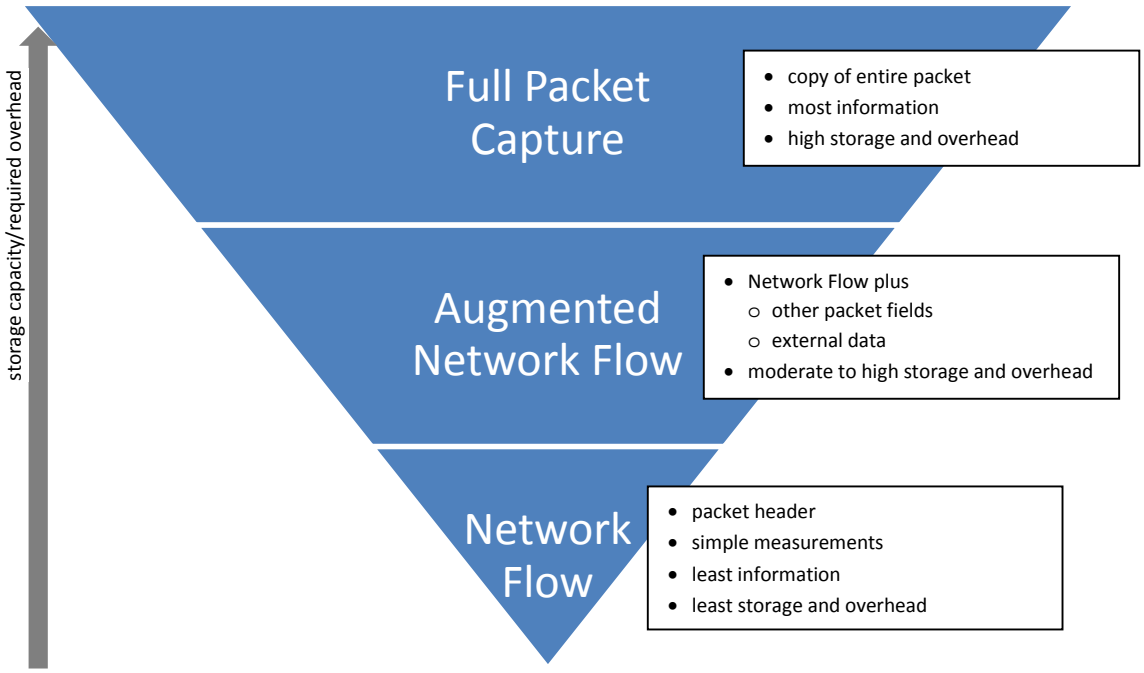

Figure 1: Network Traffic Data Tiers

1.2.1 Full Packet Capture

Full content collection, commonly referred to as *full packet capture* or *pcap*, collects the most data. This type of capture gathers and stores every packet that traverses a sensor. In other words, it collects, processes, and stores a complete copy of each traffic packet for later use. This capture data provides analysts with all header and payload information. Consequently, this tier is the most versatile for analysis at the cost of being the most resource intensive in terms of storage, processing, and analysis.

For capture of full traffic packets, the worst-case amount of raw data to be stored equals

Capacity × Time

where Capacity is the total line speed or bandwidth summed across all capture points and Time is how long the captured data must be stored

The normal-case amount of raw data to be stored is

Utilization × Time

where Utilization is the average percentage of bandwidth utilized out of total bandwidth

Using the proposed framework to do smart capturing, presented in Section 4.5, the amount of raw data to be stored equals

Utilization × Time × % Captured

where % Captured is the sum of the percentages of the total bandwidth for all categories of traffic to be kept

1.2.2 Network Flow Capture

At the opposite end of the storage spectrum is network flow. According to *RFC 3917: Requirements for IP Flow Information Export (IPFIX)* [Quittek 2004],

> *A flow is defined as a set of IP packets passing an observation point in the network during a certain time interval. All packets belonging to a particular flow have a set of common properties. Each property is defined as the result of applying a function to the values of:*
>
> 1. *one or more packet header field (e.g., destination IP address), transport header field (e.g., destination port number), or application header field (e.g., RTP header fields [RFC3550])*
> 2. *one or more characteristics of the packet itself (e.g., number of MPLS labels, etc.)*
> 3. *one or more of fields derived from packet treatment (e.g., next hop IP address, the output interface, etc.)*
>
> *A packet is defined to belong to a flow if it completely satisfies all the defined properties of the flow.*

According to this definition and the field requirements stated later in the same document, flow represents streams of network packets by generating one flow record for all packets seen for which all the following are true:

- occur within the same set time frame
- share the same source address and port
- share the same destination address and port
- use the same protocol

The record also may include some aggregate information about the flow, such as

- the total number of packets and bytes
- when the flow started and ended
- the flags seen in the different packet headers within the flow

By dropping all payloads and much header information and combining multiple packets into one record, flow decreases storage requirements. Unfortunately, it also decreases the kinds of analysis that can be performed with the information [Shimeall 2010].[1]

Though a single flow comprises a set of packets from the same source to the same destination (unidirectional), what vendors store as a flow record does vary. Some vendors store each flow in its own record, while other vendors combine two flows that they surmise to represent the conversation of two protocols into one (bidirectional) flow record. Each approach has its benefits and drawbacks.

[1] RFC 7011, *Specification of the IP Flow Information Export (IPFIX) Protocol*, a standardized network flow format, provides a more technical definition of flow. See http://tools.ietf.org/search/rfc7011.

Table 1: Comparison of Unidirectional and Bidirectional Flow

Consideration	Unidirectional	Bidirectional
Associates flows to determine conversations/sessions	Only by doing so manually	Yes
Possibility of wrongly associating two flows into a conversation or session	Only if doing manual association	Yes
Can identify when one party has more packets or bytes in the conversation/session	Yes	Only if captured and stored with the flow
Can identify who started the conversation/session	No	Only if captured and stored with the flow

Because network flow summarizes packets, the worst-case amount of raw data to be stored equals

Size × Max Flows × Time

where Size is the size of one flow record, Max Flows is the maximum possible flows per time period, and Time is how long to store the captured data

This case requires network saturation by flows of single packets with no payloads. This situation might occur in a distributed denial of service (DDoS) attack or flash crowd, but it is not likely even in those scenarios. Flow record sizes are usually a fixed length, though the actual record size varies based on the collection and storage tools.

The normal-case amount of raw data to be stored equals

Size × Average Flows × Time

where Average Flows is the average number of flows per time period

Using the proposed framework for smart capturing presented in Section 4.5, the amount of raw data to be stored equals

Size × Average Flows × Time × % Captured

where % Captured is the sum of the percentages of the total bandwidth for all categories of traffic to be kept

1.2.3 Augmented Flow Capture

Augmented flow capture encompasses everything in the spectrum between full packet capture and basic network flow. This tier adds more packet information to the flow information, either pulled directly from the header or the payload or derived from packet and flow characteristics (e.g., application labels, entropy, passive operating system fingerprinting). This tier could also contain additional information from some external source, such as the geographic location of source and destination IP addresses. Some capture solutions refer to this information as metadata because "metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" [NISO 2004]. This statement is accurate, but metadata can also refer to network packet information that has been decoupled from the traffic-defining elements of source and destination addresses, ports, protocols, and start and end times. We use the term *augmented flow* to differentiate between the solutions that keep the flow information and those that store only other types of metadata, such as only application-layer data (e.g.,

a passive domain name system [DNS] store). The information that is included in augmented flow determines if the storage requirements are closer to those of flow or full content capture.

Because augmented flow takes the network flow records of summarized packets and adds additional data, the worst-case amount of raw data to be stored equals

(Flow Size + Metadata Size) × Max Flows × Time

where Flow Size is the size of one flow record, Metadata Size is the size of the added data fields, Max Flows is the maximum possible number of flows per time period, and Time is how long the captured data must be stored

This case requires network saturation by flows of single packets with no payloads. This situation might occur in a DDoS attack or flash crowd, but it is not likely even in those scenarios. Most capture tools that support augmented flow base their records on the IPFIX standard or one of Cisco's NetFlow versions. The most common versions of NetFlow are 5 and 9, with the latter providing the most flexibility because it is template based and allows the addition of new fields to the template. The base flow record size will vary depending on what standard or proprietary format the capture solution uses.

The normal-case amount of raw data to be stored is

(Flow Size + Metadata Size) × Average Flows × Time

where Flow Size is the size of one flow record, Metadata Size is the size of the added data fields, and Average Flows is the average number of flows per time period

Using the proposed framework for smart capturing, presented in Section 4.5, the amount of raw data to be stored is

(Flow Size + Metadata Size)× Average Flows × Time × % Captured

where % Captured is the sum of the percentages of bandwidth for all categories of traffic to be kept

1.2.4 Theoretical Storage Requirements Comparison Between Tiers

Table 2 provides some example calculations for varying storage requirements between the storage tiers. We assume a 10 Gbps capacity, provisioned so that average capacity utilization is 65%. We base storage requirements for network and augmented flow on NetFlow version 9 field sizes. The nine minimum network flow fields (source IP address, destination IP address, source port, destination port, protocol, start time, end time, bytes, and packets) in NetFlow version 9 result in 33 bytes per record. The available NetFlow version 9 fields total 425 bytes per record [Powers 2010], but we use 50 bytes and 120 bytes as examples. For simplicity, we assume that all traffic is well-behaved Transmission Control Protocol (TCP) traffic with only seven packets: three handshake packets to set up the session and four handshake packets to tear it down. Because all the packets are just for session establishment and termination, we assume that all packets are the minimum TCP packet size of 40 bytes. With these givens, the calculations for storage requirements at a given tier are as follows:

Bytes per minute =

1 Gbps = 125,000,000 bytes (communication values are measured in base 10, while storage values are in base 8 [NIST n.d.])
125,000,000 × 10 = 1,250,000,000 bytes
1,250,000,000 × 60 seconds = 75,000,000,000 bytes

Full capacity per minute =

Bytes per minute / 40 bytes per packet = 1,875,000,000 TCP packets per minute
(Packets per minute / 7 packets per session) × 2 flows per session = 535,714,285 unidirectional flows per minute

Average capacity per minute =

(Bytes per minute × 0.65) / 40 bytes per packet = 1,218,750,000 TCP packets per minute
(Packets per minute / 7 packets per session) × 2 flows per session = 348,214,285 unidirectional flows per minute

Table 2: Theoretical Storage Requirements

	Size per record	Packets per record	Full capacity needed to store 1 minute of traffic over a 10 Gbps network	Average capacity needed to store 1 minute of traffic over a 10 Gbps network
Network Flow	33 bytes	3-4	16.4 GB	10.7 GB
Augmented Flow—Small Additions	50 bytes	3-4	24.9 GB	16.2 GB
Augmented Flow—Large Additions	120 bytes	3-4	59.9 GB	38.9 GB
Full Packet Capture	40 bytes	1	69.8 GB	45.4 GB

2 Literature Survey

Much of the current research focusing on network traffic capture revolves around how to capture packets at line speed with little to no data loss or how to compress the data after collection. Recent papers have discussed transforming data for efficient storage and processing [Aceto 2013a], building wire-rate capturing applications on commodity hardware [Deri 2013, Banks 2013], aggregating flow for scalable analysis [Francois 2013], and monitoring in the cloud [Aceto 2013b].

There are many standards that may provide legal requirements or best practices for network monitoring. However, they tend to be generic, often saying that some activity, such as packet capture or NSM, should be implemented, but giving little or no guidance on how to do so. A sampling of these standards includes

- Payment Card Industry Data Security Standard (PCI DSS) [PCI 2013]
- Information Security Continuous Monitoring (NIST SP 800-137) [NIST 2011]
- *ISO 27033, Information technology — Security techniques — Network security* [ISO 2009]
- The Standard of Good Practice for Information Security [Information Security Forum 2007]
- COBIT 5, *A Business Framework for the Governance and Management of Enterprise IT* [ISACA 2012]

Network flow and full packet capture do have some form of standards. Network flow formats have standards documented in RFCs, including 7011 for IPFIX [Claise 2013] and 3945 for Cisco NetFlow version 9 [Mannie 2004]. Both the IPFIX and Cisco NetFlow standards could be considered both network and augmented flow standards. Full packet capture has the loosely accepted industry standards of *libpcap* for UNIX-like systems [Wireshark 2013] and *WinPcap* for Windows systems [Riverbed Technology 2013].

Besides the literature oriented toward research and standards, there are numerous books, articles, and websites that contain how-to information, both generic and geared toward individual products. The books *The Practice of Network Security Monitoring Understanding Incident Detection and Response* [Bejtlich 2013] and *Applied Network Security Monitoring: Collection, Detection, and Analysis* [Sanders 2013] introduce NSM topics, including traffic collection, by using open source tools to illustrate concepts. The paper *Logging and Monitoring to Detect Network Intrusions and Compliance Violations in the Environment* [Gupta 2012] and the article "Network Monitoring as a Security Tool" [Moyle 2012] both discuss using network monitoring for security purposes. Vendor websites are a good starting point for information on individual products; sites like TechTarget[2] and Dark Reading[3] have many articles related to security monitoring as well as vendor and book reviews.

[2] http://www.techtarget.com/

[3] http://www.darkreading.com/

3 Solution Considerations

Organizations looking for a capture solution have many vendors to choose from, and many vendors offer multiple products. When evaluating a solution, there are many aspects to consider such as price, hardware and software requirements, and scalability. Of specific interest for the scope of this report are the aspects related to storage. In this report, it is not possible to evaluate all vendor offerings or cover all possible considerations, nor do we intend to provide specific vendor product suggestions. This section highlights the main storage-related considerations to consider when evaluating and selecting a capture solution from any vendor. Any mention of a specific vendor or product is for example purposes only and is not to be construed as a recommendation or endorsement.

The storage-related aspects of a capture solution fall into two categories: filtering what to store and the storage method. *Filtering what to store* relates to a solution's ability to let organizations choose what network traffic is important and what the solution should store for later analysis. The *storage method* is how the solution packages the captured data and where it is kept.

3.1 Filtering

Filtering enables organizations to choose what network traffic they want the capture solution to store. When evaluating a solution's filtering capabilities, it is important to look at what filtering is available and where the filtering occurs.

The ability to filter has several aspects. A solution may have predefined filters or provide the ability to create custom filters. A solution's filter may be able to select traffic based on its characteristics at the network level or the application level. Network-level characteristics include IP address, port number, and packet size. Application-level characteristics include the application that generated the traffic. Some solutions offer a special type of filtering by sampling instead of or in addition to other filtering capabilities. Sampling chooses packets to keep based on an algorithm, such as keeping every 10th packet or choosing packets pseudo-randomly. Some capture solutions also allow the user to configure what portion of the interesting traffic should be stored. For instance, Berkeley Packet Filter (BPF) enables users to save the first *x* number of bytes of each network packet that passes a filter [McCanne 1992]. To effectively implement the smart collection and storage method explained in Section 4, a solution must, at a minimum, allow customizable filtering based on network-level characteristics, though filtering on application-level characteristics is even better.

Filtering impacts performance and can change the storage requirements of the system, depending on which type of filtering is implemented: in-line or post-processing. *In-line filtering* takes place during the actual traffic capture process. *Post-process filtering* takes place after initially writing captured traffic data to disk.

In-line filtering can be hardware-implemented using special network capturing cards, as with the Napatech NT40E2-1 network adaptor [Napatech 2014], or can be software-implemented, as with a Berkeley Packet Filter [McCanne 1992]. Either way, filtering occurs before the solution puts the traffic data in the buffer queue for writing to disk. In-line filtering is good for storage because on-

ly the desired packets are ever written to disk, but it increases processing requirements. Consequently, filtering decreases how much data can be captured per unit of time, meaning that filtering decreases the line speed a solution can handle; the greater the complexity of the filter, the lower the manageable line speed. In general, hardware implementations can handle greater line speeds than software implementations, but both are limited.

Post-process filtering is software-implemented and is often a manual process. This type of filtering reads the saved data from disk and chooses which data to save. For instance, a Wireshark user can open a data file, apply a filter, and save the results. Some solutions may enable the user to create an automated method that executes post-process filtering, if the solution provides a command-line interface or application programming interface (API). This type of filtering requires more short-term storage capacity than in-line filtering does, even when the filtering results are the same traffic to be stored long-term.

Either in-line or post-process filtering will work for the smart collection and storage method described in Section 4. However, the calculations presented there assume in-line processing. If that is not available, organizations must consider their storage needs for keeping data before the captured traffic can be filtered. The best option is a solution that offers both in-line and post-process filtering.

Table 3: Filtering Considerations

Consideration	Available
Predefined Filtering	[ ] Y [ ] N
Customizable Filtering	[ ] Y [ ] N
Filtering on Network-Level Characteristics	[ ] Y [ ] N
Filtering on Application-Level Characteristics	[ ] Y [ ] N
Configuration for Saving Specific Byte Count of Filtered Packets	[ ] Y [ ] N
Sampling	[ ] Y [ ] N
In-line Filtering	[ ] Y [ ] N
Post-Processing Filtering, Manual	[ ] Y [ ] N
Post-Processing Filtering, Automated	[ ] Y [ ] N
Line Speed Supported	_____

3.2 Storage Methods

Collected data must persist long enough for analysts to consume what they need from it. How and where the data is stored can dramatically impact the total storage capacity required and other storage-related costs (e.g., administration and upkeep). Organizations may choose to keep the data locally or in a cloud or other external storage. Furthermore, whether keeping the data local or moving it to a cloud, organizations will often need to decide between using file-based storage, a database, or a combination. Each has its own considerations. In this section we first discuss considerations for keeping data locally, then considerations for cloud storage. We conclude with a discussion about file-based and database storage.

3.2.1 Storage within the Organization

Most organizations have network traffic capture solutions that collect data at multiple points in the network. Therefore, the physical location where collected data is stored is important. Storing

all collected data in one central physical location may simplify management and analysis of the data, but it requires moving data to the central location, usually across the network, which consumes network bandwidth. This also concentrates the data, making exfiltration simpler if the repository is compromised. Alternatively, keeping the data where it was collected does not regularly consume network bandwidth. However, using this kind of distributed storage may increase analysis complexity, especially if analysts must physically log into several storage locations to access the data they need.

The main consideration for keeping data within the organization is the storage media's (e.g., network-attached storage, external hard drives) purchase and upkeep costs. Organizations must also keep traffic captures secure, treating them as they would other sensitive information.

3.2.2 Cloud Storage

Cloud storage currently comes in two flavors: (1) storage that is bundled with the capture solution and is administered by the solution vendor and (2) storage purchased from a vendor not associated with the actual capture solution. Either way, there are a number of important considerations for organizations evaluating cloud storage.

Cloud storage replaces much of the costs for storage media and upkeep with daily pay-as-you-use fees, but it has some additional costs. When comparing costs between keeping data in-house and moving it to a cloud, organizations must consider the additional costs for cloud storage:

- transporting the data to the storage location
- transporting the data from storage when needed for analysis
- securing the data during transport and while stored
- conducting analysis
- managing data retention
- training and administration

In addition to evaluating the monetary costs, organizations must also evaluate the technical feasibility of storing and retrieving cloud-based data for use, whether the level of service (i.e., availability) meets analysts' needs, available recourse if the cloud provider ceases business, and if the data can be secured sufficiently to meet the organization's policies and applicable laws. Organizations must also identify any data specified by law that must be stored within the United States or have its physical location tracked.

3.2.3 File-Based Storage

Most capture solutions write out the captured data to files. Organizations need to consider the properties of the stored files, such as format and overhead, if compression methods can be applied, and, if so, which ones.

3.2.3.1 Format and Storage Overhead

When evaluating solutions that store data in files, it is important to know what file formats the solution supports. For transferring data between applications or organizations, a solution that supports only proprietary file formats may not be a good choice. There are some common formats,

such as .pcap, that many capture-and-analysis applications can handle, so a solution that saves as that format or can export to that format may be a better option.

When determining storage capacity requirements for a solution, the file formats supported and the vendor's implementation of the formats can influence the required amount of storage. While for a small volume of data the differences between formats may seem small—usually around 1%—the required amount of storage adds up quickly.

Table 4 illustrates this point, using the storage required for one day when traffic is at 1 Gbps, 10 Gbps, and 100 Gbps and file size is set to store 50 MB of traffic data. First, we translate network speed to disk storage requirements:

1 Gbps = 1,000 megabits per second

1,000 megabits / 8 bits per byte = 125 MB of hard drive storage per second

125 MB × 86,400 seconds per day = 10,800,000 MB a day

10,800,000 / 50 MB files = 216,000 files needed for one day of network data

This data fills out the columns in the table:

- Disk Space for 50 MB Raw Data = actual file size for a vendor's implementation of a file format storing 50 MB of raw network data
- % Increase = additional storage percentage above Wireshark's implementation of .pcap, chosen because it is the smallest total file size of those tested
- Additional Storage at 216,000 Files = additional storage volume for one day's worth of network data at 1 Gbps above Wireshark's implementation of .pcap (additional requirements are in TB, not MB)
- Additional Storage at 2,160,000 Files = additional storage volume for one day's worth of network data at 100 Gbps above Wireshark's implementation of .pcap
- Additional Storage at 21,600,000 Files = additional storage volume for one day's worth of network data at 100 Gbps above Wireshark's implementation of .pcap

Table 4: File Format Savings Example

Vendor	Extension	Disk Space for 50 MB Raw Data	% Increase	Additional Storage at 216,000 Files	Additional Storage at 2,160,000 Files	Additional Storage at 21,600,000 Files
Wireshark	.pcap	51,201 KB	-	-	-	-
Wireshark	.pcapng	51,959 KB	1.01%	156.14 TB	1,561.43 TB	15,614.32 TB
Nokia	.pcap	51,369 KB	1.00%	34.61 TB	346.07 TB	3,460.69 TB
RedHat 6.1	.pcap	51,538 KB	1.01%	69.42 TB	694.20 TB	6,941.99 TB
SuSE 6.3	.pcap	51,706 KB	1.01%	104.03 TB	1,040.27 TB	10,402.68 TB
HP-UX nettl	.trc0	53,391 KB	1.04%	451.13 TB	4,511.26 TB	45,112.61 TB
Microsoft NetMon 2.x	.cap	51,369 KB	1.00%	34.61 TB	346.07 TB	3,460.69 TB
Accellent 5View	.5vw	52,043 KB	1.02%	173.45 TB	1,734.47 TB	17,344.67 TB
TamoSoft CommView	.ncf	51,622 KB	1.01%	86.72 TB	867.23 TB	8,672.33 TB
K12text	.txt	154,781 KB	3.02%	21,336.82 TB	213,368.23 TB	2,133,682.25 TB

It is important to consider these differences when calculating storage requirements for full packet capture. The calculation of overhead per file for each file format is

File Size – Raw Data

For the example in Table 4, we know that 50 MB of raw data is 51,200 KB. Using Wireshark's .pcapng values from the table, we can calculate the overhead: 51,959 KB – 51,200 KB = 759 KB per file.

3.2.3.2 Compression

Data compression can make storage and analysis more efficient. A good compression algorithm may not only lessen the disk space required to store data, but also improve the time it takes to retrieve that data from disk. As an example of storage savings, SiLK, from the CERT® NetSA Security Suite, has an uncompressed, fixed record size of 52 bytes. It supports two compression algorithms, lzo1x and zlib. The lzo1x algorithm reduces the record size by about 50% [SEI 2014]. For file storage, if a capture solution does not itself compress data, organizations may be able to compress it themselves, such as with gzip. However, they will need to uncompress the files before analysts can use them, and organizations will need to consider the storage needed to hold uncompressed files while they are under analysis.

Table 5: Storage Method Consideration

Consideration	Available
File-Based Storage	[ ] Y [ ] N
Database-Based Storage	[ ] Y [ ] N
Central Storage Supported	[ ] Y [ ] N
Distributed Storage Supported	[ ] Y [ ] N
Organization Responsible for Storage	[ ] Y [ ] N
Vendor Responsible for Storage	[ ] Y [ ] N
Multiple File Formats	[ ] Y [ ] N
Data Compression	[ ] Y [ ] N
Data Compression Ratio	_____

3.2.4 Database Storage

Some capture solutions can be configured to write to a database. For solutions that support only files, it is possible to store the files themselves in a database, either with a third-party tool, a manual process, or a scheduled task. Organizations need to consider the how the selected method scales and what security can be implemented.

Furthermore, when evaluating solutions that store data in a database, organizations must consider the supported databases and special requirements the solution needs in order to work with them. They also need to evaluate the expected size and record count of their storage volume and be aware of any database limitations restricting total database size, record size, or record counts. Relational databases are not as scalable as file-based storage, so if the solution uses a relational database (as opposed to a NoSQL database like Hadoop), organizations must ensure that scalability is

® CERT is a registered mark of Carnegie Mellon University.

not an issue. Organizations should also realize that, in general, database storage methods have more overhead for raw traffic than most file format storage methods.

4 Smart Collection and Storage

The smart collection techniques discussed in this report fall into two categories: tiered storage and protocol/service-level filtering. In this report, we use the term *tiered storage* to refer to the concept of choosing to store either whole network packets or different parts of network packets, based on some specified criteria. *Protocol/service-level filtering* refers to the concept of completely disregarding network packets if they are of a specific protocol or service.

4.1 Methodology

To determine the optimal storage policy, organizations follow two broad steps:

1. Quantify the risks and effectiveness of storing each type of network traffic at each tier.
2. Evaluate how those values change over the time of storage.

If desired, organizations can then combine the values with static costs, such as software and hardware, into a cost-benefit analysis.

In this section, we first present considerations specific to each component—capture tier, traffic, and time. We describe a method for quantifying the risks and effectiveness of captures and then discuss using the results to formulate a storage policy that best meets an organization's needs within its limitations. Finally, we present how to use the policy to plan for growth, which is an important part of implementing any capture solution.

In Section 5, we provide examples that show how smart collection can benefit several real-world networks.

4.2 Capture-Specific Considerations

Understanding the benefits, limitations, and risks of the different storage tiers is necessary to make knowledgeable decisions about what data to store and where. This section provides some considerations specific to each of the three tiers. It is important to secure all tiers. If the data is subject to unauthorized access, it might be corrupted, deleted, or exfiltrated and become unusable to analysts or a security threat to the organization—important network information can be inferred from data at all tiers, such as IP addresses of internal hosts and what services are available on different IP addresses.

4.2.1 Full Packet Capture

As already mentioned, full packet capture has the most intense storage requirements of the tiers. This type of capture also has the most intense analysis requirements, both because analysts have many more records to examine pertaining to any given problem and because those records contain extraneous information beyond the data they need. The sheer volume of data may lead to high retrieval costs in a given analysis tool, including long retrieval times and high processor utilization, to find the data that matches a query. Organizations should be aware that though full packet capture provides the most information to analysts, accessing that information can be burdensome, in terms of both analyst labor and network overhead. Storing the same data at multiple tiers,

though requiring extra storage, may be necessary to allow analysts to efficiently and effectively remediate incidents and perform their other duties.

Besides storage and analysis complexity, the contents of full packet capture also merit consideration. Storing all the data that traverses a network entails a high probability of the contents containing sensitive but unclassified (SBU) data, such as personally identifiable information (PII), that may be subject to laws requiring its protection, and there is no guarantee that such data is encrypted. Besides protected SBU data, the traffic contents are also likely to contain information such as intellectual property. Some of the captured information could make the traffic capture storage subject to e-discovery rules. Because it is possible to retrieve detailed information with access to the data, it is very important to ensure that full packet captures are securely processed and stored. It is also a good idea to check with legal counsel on what full content information may not be or must be stored.

4.2.2 Augmented Flow

Augmented flow storage requirements vary widely based on the information used to augment network flows. Depending on the number of fields the solution retains and whether files from the traffic sessions are extracted and stored, the storage requirements for augmented flow could come close to or even exceed those for full packet capture. Organizations should monitor how the record size changes as they select augmented flow fields for their solution. If they plan to extract and store files, they also need to determine how much extracted files will add to the storage requirements. If the organization relies on metrics based on trend analysis, changing the capture parameters may also affect the validity of those metrics.

It is unlikely that stored, strictly augmented flow will contain PII or business confidential information. However, organizations should treat extracted and stored files the same way they would full packet captures.

4.2.3 Network Flow

Because network flow data is the most compact form of capture, each gigabyte of storage capacity can retain more network flow records than augmented flow records or full packets. This means that as new network flow records enter storage, the older records can be retained for longer, which lengthens the time frame available for analysis. This type of capture benefits analysis that takes place a long time after an activity has occurred, such as an investigation of advanced persistent threats. Unfortunately, the minimal information available with network flow means some important aspects of activities cannot be investigated, such as what files were exfiltrated.

4.3 Traffic-Protocol-Specific Considerations

Each category of traffic protocol has features that determine how it can be used and where it may or may not be useful. Organizations should consider these characteristics when determining the importance of their network traffic. It is also beneficial to determine if other existing applications or products already store the data. For instance, email is often already stored long term on an exchange server, and organizations may store VoIP call log information. Duplicating this data provides little benefit. This section presents a few common considerations for storing and analyzing different types of traffic; however, this discussion is not comprehensive. As part of determining

how to configure smart data collection, the organization should give each of its defined traffic types careful thought and consideration.

4.3.1 Encrypted Protocols

Encrypted traffic, such as HTTPS, presents a challenge to analysts. If the traffic passes the capture sensor in an encrypted state, analysts, in most cases, cannot examine the packet contents, even with full packet capture. Organizations should consider if it is worth storing packet content when the packet is in a state that analysts cannot analyze and if an augmented flow solution can provide more information than network flow for encrypted packets. If organizations need to be able to analyze SSL certificates, it may be possible to capture them with augmented flow or just to store the first several packets that contain the certificates in full packet capture.

4.3.2 Email

Email traffic can be a large portion of an organization's network traffic. Many organizations are subject to laws, such as those for e-discovery, or industry regulations requiring that email be stored for a set period of time. Organizations usually fulfill these requirements by storing all sent and received email on a mail server. For organizations that already store emails, capturing the content of email traffic would result in duplicate storage. The email traffic header or flow information may prove valuable, so organizations should consider what information can already be obtained from the actual emails that are already being stored and what, if any, additional information would be helpful to capture.

4.3.3 Domain Name System

DNS traffic occurs frequently, as devices look up IP addresses for domain names when users access websites or devices check for software updates. This information can be immensely valuable for network defense and incident response. Organizations may find that a dedicated DNS capture solution may fit their needs better than using a generic packet capture solution, especially because DNS information does not package well into network or augmented flow. Organizations that have a passive DNS server may also find that much of the information they need related to DNS is already available in their passive DNS database and that packet capture provides little added benefit.

If organizations decide that capturing DNS traffic with their generic solution provides benefits, they should consider if there are differences in usefulness between Transmission Control Protocol (TCP) DNS traffic and User Datagram Protocol (UDP) DNS traffic. Most DNS is UDP. However, DNS uses TCP if an answer to a DNS query is larger than that allowed by UDP, as well as for some DNS administrative functions, such as zone transfers, that can consume much bandwidth.

4.3.4 Voice Over IP

Internet phone calls, often referred to as voice over IP, or VoIP, may present legal issues. This traffic is a form of telephony; hence, wiretap and phone call recording laws may apply to full packet traffic capture. This issue should be discussed with legal counsel to determine if it may be an issue for the organization.

VoIP also has a lot of overhead. VoIP signaling protocols are active on devices running VoIP applications. These applications could be used to detect what devices run VoIP but may not provide

much other benefit. Organizations can typically retrieve call information, such as call length or who called whom, from the VoIP server, where it is kept for billing and auditing purposes.

4.3.5 Network Time Protocol

The network time protocol (NTP) is a service used to synchronize the clocks of devices on a network. While in general this traffic may be a very small portion of capture traffic, this protocol is vulnerable to exploitation for use in distributed reflected denial of service attacks if organizations do not appropriately secure their NTP servers. These attacks may generate hundreds of gigabits of traffic per second.

4.4 Considerations for Time to Store

When determining an appropriate length of time to store traffic capture data, it is important that organizations honestly evaluate

- how long the data is useful
- how long until the data becomes useful

Old data is not always irrelevant, and new data is not necessarily the most valuable. This is especially true when organizations do not detect incidents until months after the fact, which is the norm [Verizon 2013]. Data needs to be available when needed. If organizations regularly discard data before it has ever been useful, they should reevaluate their processes and determine if the data should even be stored in the first place. Organizations should also ensure they store traffic that is subject to service level agreements or legal obligations for the required periods of time.

4.5 Criticality and Effectiveness Ranking

Determining what traffic to capture and store is not a simple task. Not having data available for analysis can have negative financial and productivity impacts. Ideally, organizations would store and analyze all traffic that traverses their networks, but as mentioned earlier, this is not practical for multiple reasons. Consequently, it is important for organizations to capture and store traffic in a format that provides the best effectiveness within the organization's limitations. This section outlines a method for ranking the criticality and effectiveness of different types of traffic. These rankings will determine the tier at which the organization will store each traffic type.

Each organization must determine its own criticality and effectiveness values, which are highly dependent on the individual organization's needs, assets, abilities, and regulatory requirements. The process for completing the method described in this section should be a group effort, not a task done by one individual within the organization. Input should be gathered from the analysts who need the data for incident response or forensic investigations, network administrators who need the data for monitoring network growth and trends, managers who set budgetary constraints, and others versed in the legal implications of storing or not storing specific types of data.

The method presented uses a series of charts to guide organizations to a filtering and storage policy that yields the highest value while staying within storage, budget, and other constraints. There are six steps to filling out the five charts. It is assumed that the organization already knows baseline network usage patterns and trends.

Step 1 identifies applicable reasons for using captured data.

Step 2 starts the process of determining how long it is desirable to store the captured data.

Step 3 identifies the attack categories, and the data necessary to investigate them, that are most relevant to the organization.

Step 4 finds the storage time frame that is most effective for storing data related to a specific attack category.

Step 5 ranks how essential different categories of traffic are to the organization's investigation needs.

Step 6 calculates storage requirements and guides selection of the traffic to store at each storage tier.

Final take-aways from the process are

- what traffic to filter out of each tier
- the total storage requirements necessary to store each tier for an appropriate time period

Step 1: Network Data Usage

Step 1 employs the *Network Data Usage Chart* (Table 6) to identify applicable reasons for using captured data. Different uses for the data require different levels of information, as shown in the last three columns of the table. The "Purpose" column lists the possible uses for traffic data, and the following columns show what the different tiers of data can tell about the traffic for the given use. Select all possible reasons why your organization would ever use network traffic data.

Table 6: Network Data Usage Chart

Y/N	Reason Number	Purpose	Capture Method*	Who	How Much	When	How Long	Using What	Transfer-ring What	How
	1	Investigate attacks	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓	✓	✓	
__			P	✓	✓	✓	✓	✓	✓	✓
	2	Police policies	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓	✓	✓	
__			P	✓	✓	✓	✓	✓	✓	✓
	3	Provide information to create policies	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓	✓	✓	
__			P	✓	✓	✓	✓	✓	✓	✓
	4	Understand normal network traffic	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓			
__			P	✓	✓	✓	✓			
	5	Understand abnormal network traffic	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓			
__			P	✓	✓	✓	✓			
	6	Plan for network upgrades	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓			
__			P	✓	✓	✓	✓			

*** N = network flow, A = augmented flow, P = pcap**

Step 2: Number of Storage Days

Step 2 uses the *Number of Storage Days Chart* (Table 7) to start the process of determining how long it is desirable to store the captured data. It also encourages the organization to think about how long traffic data at different tiers is useful for different purposes. Judging how long data is useful is, for most of the purposes, straightforward. Using data to investigate or respond to attacks is the exception. For Step 2, begin to fill in the *Number of Storage Days* chart by entering the number of "Days of Traffic Needed" for each of the purposes selected in the *Network Data Usage* chart. Leave Purpose 1 blank for all three capture methods; these values will be determined later in this methodology.

Table 7: Number of Storage Days Chart

Capture Method	Selected Network Data Usage Purpose	Days of Traffic Needed	Notes
Pcap	1	To investigate attacks: ___	See Attack/Risk and Requirements Charts
	2	To police policies based on traffic content: ___	Consider using filtering/IDS instead
	3	To see what content is passing to create network use policies: ___	Consider using filtering/IDS instead
Augmented Flow	1	To investigate attacks: ___	See Attack/Risk and Requirements Charts
	2	To police policies based on traffic content: ___	Consider using filtering/IDS instead
	3	To see what content is passing to create network use policies: ___	Consider using filtering/IDS instead
Network Flow	1	To investigate attacks: ___	See Attack/Risk and Requirements Charts
	2	To see volume, who, and when information to police policies based on traffic content after the fact: ___	
	3	To see volume, who, and when information to create network use policies: ___	
	4, 5	To trend traffic to understand normal and abnormal traffic: ___	
	6	To trend traffic to plan for upgrades: ___	

Step 3: Attack Type Criticality

Steps 3-6 help organizations determine what traffic they need to store and for how long in order to support their investigation of or response to attacks. Step 3 starts the process by identifying the attack categories that are most relevant to the organization and the data necessary to investigate them.

How organizations view different attack categories is important in determining the traffic that is most necessary to capture and how long to store that traffic. This criticality is organization specific and should be determined for each type of attack category an organization may experience. Or-

ganizations must base the criticality on what their analysts need to know to investigate or respond to the attack. In the *Attack Type Criticality Chart* (Table 8), organizations should enter a criticality value for each row and column. Determine the criticality based on the likely impact if investigation of an incident of a specific category (row labels) could not occur without a particular type of information (column labels). Table 8 is an example of one possible categorization of attacks. Organizations can change the attack categories to fit their own needs. See Appendix A for a more detailed example of this chart and the other ranking charts.

Table 8: Attack Type Criticality Chart

		Criticality of determining:		
		Network Flow (N)	**Augmented Flow (A)**	**Pcap (P)**
		who is affected, how much, when, how often	**using what, transferring what**	**how and what data**
Attack Category	**Reconnaissance**			
	Exploit			
	Exfiltration			
	DoS/DDoS			

Criticality Levels
3: Legal ramifications/major increase in financial losses
2: Increase in financial losses or time to resolve issues caused by the incident
1: None or minor inconvenience

Step 4: Effectiveness

Step 4 uses the *Effectiveness Chart* (Table 9) to find the storage time frame that is most effective for storing data related to a specific attack category. Organizations define the time frames as a set number of days. The effectiveness ranking quantifies the best time frame during which having the data is most effective relative to the other time frames. At some point in time, the added value of keeping the data longer is negligible; the incident may no longer be relevant (e.g., a "statute of limitations"), or logistical reasons may prevent analysts from accessing data beyond the time frame (e.g., there is just too much data to look through, or beyond some point the data is archived and difficult to access). Effectiveness can strictly increase over time, strictly decrease over time, increase for a period of time then decrease, or remain steady. When filling out this chart, effectiveness should never decrease then increase.

In the *Effectiveness Chart* (Table 9), first define the storage time ranges for immediate, short-term, mid-term, long-term, and archive effectiveness (columns 2-6) as a set number of days. Enter the upper value of each range as the column value. Then enter an effectiveness value for each attack category at each time range. Organizations should change the attack categories used here to match the categories they entered in the *Attack Type Criticality Chart* (Table 8).

In each attack category's "Days of Greatest Effectiveness" cell, enter the column value for the effectiveness time frame with the highest effectiveness value for the attack category. Again, row values should trend up, trend down, or be steady—there should not be multiple peaks. If multiple

effectiveness time frames tie for the highest effectiveness value, enter the greatest column value that corresponds to the longest time frame.

Table 9: Effectiveness Chart

	Column Value:	____	____	____	____	____	
	Effectiveness Time Range (in days):	**Immediate**	**Short-term**	**Mid-term**	**Long-term**	**Archive**	**Days of Greatest Effectiveness**
Attack Category	**Reconnais-sance**						
	Exploit						
	Exfiltration						
	DoS/DDoS						

Effectiveness Value
5: Five times the effectiveness of value 1—data at this time frame is critical
4: Four times the effectiveness of value 1
3: Three times the effectiveness of value 1
2: Two times the effectiveness of value 1
1: Data at this time frame has little use

Step 5: Essentiality

Step 5 uses the *Risk and Requirements Chart* (Table 10) to capture the relevance and relative effectiveness of different types of traffic that it is possible to capture with a traffic capture solution. Organizations can then use this information to calculate storage requirements based on how long they should keep the data in each tier of storage. Organizations can divide the traffic into types or services in whatever manner they desire; we will refer to these as *services*. The method must produce categories that have no overlap, are measurable, and can filter out traffic in the organization's capture solution. Section 5.1 shows one possible method.

Step 5a: Service and Attack Type Criticality

Just as in the *Effectiveness Chart*, change the attack categories of the *Risk and Requirements Chart* to match those used in the *Attack Type Criticality Chart*. Evaluate analysts' ability to detect or investigate each combination of traffic type or service, attack category, and storage tier. Black out the cells for combinations that are unlikely to be useful. In each remaining cell in the "Attack Category/Storage Tier" columns, enter the corresponding criticality value from the *Attack Type Criticality Chart*.

Step 5b: Traffic Storage Risks

In the "Risk Value" column of the *Risk and Requirements Chart*, enter a risk value for storing each service at each storage tier. Determine risk by balancing criticality against possible ramifications of storage or data leakage. The *Storage Tier Risks Chart* (Table 11) lists common risks associated with each tier of storage. Each organization should evaluate these risks against the laws and standards that apply to the organization. In general, the risk associated with storing network flow data is possible leakage of a network's structure and traffic patterns if the capture solution storage is compromised. All organizations should consider this the baseline risk.

Step 5c: Traffic Essentiality

In the "Essentiality" column of the *Risk and Requirements Chart*, calculate a relative value for how essential each service is to the organization at each tier. Do the following for each service:

1. Identify the network flow and attack category combination with the greatest criticality (refer to the *Attack Type Criticality Chart*). Find the corresponding combination in the *Risk and Requirements Chart*.
2. Subtract the network flow risk value of the corresponding combination identified in step 1 from that combination's criticality value (also determined in step 1).
3. Enter the result of step 2 in the "Essentiality" column's network flow cell for the service.
4. Repeat steps 1 through 3 for augmented flow and full packet capture.

To eliminate negative values, add 4 to every result if using a 3-point criticality scale and a 5-point risk scale. When using other scales, add the absolute value of the lowest possible result of subtracting risk from criticality.

Step 5d: Useful Days

Determine what attacks can be detected for each service (i.e., any attack whose cells for the service are not blacked out). Of the attack categories that are detectable for a given service, find the attack category with the highest value in the "Days of Greatest Effectiveness" column from the *Effectiveness Chart*. Enter that value in the "Useful Days" cell of the "Effectiveness" column for the service.

Step 6: Storage Requirements

Step 6 uses the *Risk and Requirements Chart* (Table 10) to help calculate storage requirements and to guide the selection of the network traffic to store at each storage tier. This selection requires traffic measurements of the organization's network.

Step 6a: Measuring Traffic Volumes

Measure traffic for the previous 9-12 months, capturing both weekdays and weekends. The 9-12-month time frame should encompass enough data to smooth out nonrepresentative traffic volatility typical of shorter periods. For instance, as shown in Figure 2, traffic volumes at DE-CIX, a German internet exchange,

- grow a little in January and February
- are fairly steady March through June
- dip in July and August
- grow rapidly the rest of the year

Consequently, measuring a short time period with a downward trend in traffic can cause the organization to underestimate its needs, especially when planning for needs several months down the road. The reason it is important to capture both weekdays and weekends in the measurements is that some organizations may see higher traffic outside normal business hours. This traffic could be caused by administrative activities, such as backups, or by general internet users accessing the organization's websites.

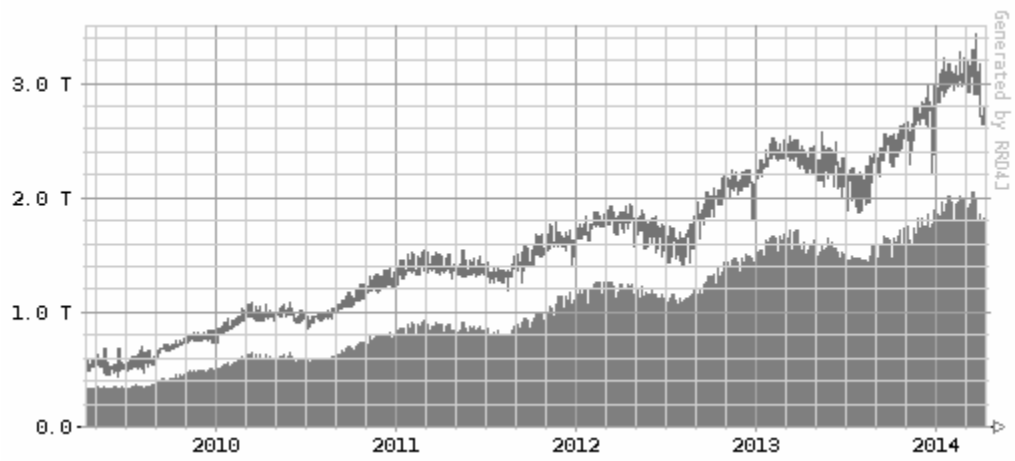

Figure 2: Long-Term DE-CIX Frankfurt Traffic Volumes [DE-CIX 2014]

The minimum statistics captured during measurement should include the bytes for each service type per day and number of flows for each service type per day. Organizations should use these values to calculate the current expected traffic volumes per day and note any trends in volume changes.

How organizations calculate expected traffic volume can vary, but the method used must be accurate or an overestimate. One method for calculating expected volume for each service is to use each service's single greatest volume from the entire measurement period. In other words, total each service's bytes for each day of the measurement period and choose the day that has the highest value. Repeat for the number of flows for that service. Do these two steps for each service in the chart. This method ensures that storage meets foreseeable maximums for the current traffic patterns. At this step, it is not important to account for future traffic volume increases; Step 6d covers projected growth. Organizations should enter the expected volume values for each service in the "Expected Traffic per Day" column for both bytes and flows.

Step 6b: Storage Requirements per Service

Once the daily expected traffic volumes for each service have been determined, calculate how much storage it would take to store the data at each tier. These values go in the "Storage per Day" column. The calculations for each tier are

$$N = \text{Max Flows} \times \text{Network Flow Record Size}$$

$$A = \text{Max Flows} \times \text{Augmented Flow Record Size}$$

$$P = ((\text{Bytes} / \text{File Size}) \times \text{Overhead}) + \text{Bytes}$$

In the network and augmented flow equations, *Flows* is the value from the "Max Expected Traffic per Day" column's "Flows" cell for the service. *Network Flow Record Size* and *Augmented Flow Record Size* are the number of bytes one network or augmented flow entry requires for storage. This number of bytes depends on the capture solution and what fields it is set to store. *Bytes* is the value from the "Max Expected Traffic per Day" column's "Bytes" cell for the service. In the full packet capture equation, *file size* is how much raw data to store per file, and *overhead* is the overhead for the file based on the storage format as explained in Section 3.2.

Step 6c: Where to Store Each Service

Organizations can now determine at which tier each type of traffic should be stored. Organizations should base their choices on the storage requirements, criticality, and effectiveness value for each service. At this point, assume that traffic at higher tiers is also stored at the lower tiers. Once the organization has determined where to store each service, it should sum the values of the applicable services for each tier to get the total daily storage requirements. For augmented flow, add extra overall storage to the totals for extracted files if the organization will be extracting files and storing them as part of its augmented flow capture solution.

Step 6d: Total Storage Requirements

This step determines the total storage requirements that organizations need to determine the historical time frame for each tier of data. Of the services that will be stored in network flow, find the greatest useful days value and enter it in the "Storage Days Goal" network flow cell. Repeat for augmented flow and full packet capture. For network flow, check the *Number of Storage Days* chart, and if any day value in the chart is greater than the "Storage Days Goal" for the network flow column, change the "Storage Days Goal" to the greater value. "Storage Days Goal" values are the target number of days to keep traffic data at each tier.

If network traffic is growing for the organization, it must calculate projected storage requirements for one day's worth of data for the day before it increases storage. In other words, if the organization will not increase storage for two years, it needs to calculate how much data will need to be stored for the last day of the two-year period. Organizations can make this calculation by applying the traffic volume growth trends noted in Step 6a to the current values using the formula

$$\text{Current Storage} \times ((1 + \text{Growth Rate})^{\text{Months}})$$

Current storage is the storage required for one day's worth of data from the *Risk and Requirements Chart* at the current levels of network traffic. *Growth rate* is the monthly rate of growth, as a percentage, found from volume trending. *Months* is the number of months that the storage must be sufficient. Organizations may apply this formula to the totals for each storage tier using the average volume growth rate, though applying the formula to each service would provide a more accurate estimate. If using this second option, calculate growth for each service and apply it to each service that will be stored at each of the tiers. Then multiply the projected totals by the "Storage Days Goal" values to obtain the total storage requirements for each storage tier.

If the resulting values are unattainable for any tier, use the criticality and effectiveness values to prioritize traffic types and storage days. Organizations can do this in several ways depending on their resources and capabilities. One way is to start moving traffic types to storage only in a lower tier, starting with those having the lowest effectiveness values. Another way is to move storage closer to the goal by applying filtering to a tier after some time to discard less effective traffic services while continuing to retain the most important services. Beyond helping to prioritize what traffic types to store, the criticality and effectiveness values can also be used to justify increased storage funding.

Once all five charts are completed, organizations know what traffic they will need to store as network flow, augmented network flow, and full packet capture. They will also know how much storage capacity is currently required for one day's worth of data at each tier, as well as how much

storage they need to store the necessary data until they can next increase their storage or update their system.

Table 10: Risk and Requirements Chart

		Attack Category/Storage Tier												Risk Value			Essentiality			Effectiveness	Max Expected Traffic per Day		Storage per Day		
		Reconnais-sance			Exploit			Exfiltration			DoS/DDoS														
		N	A	P	N	A	P	N	A	P	N	A	P	N	A	P	N	A	P	Useful days	Bytes	Flows	N	A	P
Traffic Type/Service	TCP													1											
	UDP													1											
	ICMP													1	1	1									
	OTHER													1											
																						Totals:			
																						Storage Days Goal:			

N = network flow, A = augmented flow, P = pcap

Risk Values

5: Storage of information has legal ramifications that cannot be mitigated.

4: Storage or leakage of information has legal or major financial ramifications that can be mitigated, but mitigation is not already in place.

3: Storage or leakage of information has financial ramifications that can be mitigated, but mitigation is not already in place.

2: Storage or leakage of information has legal or financial ramifications, but mitigations are in place.

1: Storage or leakage of information would cause minor inconvenience. This value is used anywhere the ramifications are unlikely to be anything other than leaked network structure if the stored data was subject to unauthorized access.

Table 11: Storage Tier Risks

Tier	Description
Network Flow	• Network structure
Augmented Flow	• PII • E-discovery (if storing embedded files only) • Malware (if storing embedded files only) • Business confidential data • Network structure
Pcap	• PII • E-discovery • Malware • Phone conversations (TCP: VoIP only) • Web cam content • Credentials (from unencrypted traffic) • Business confidential data • Network structure

4.6 Planning for Growth

When organizations purchase and implement a capture solution, they must plan for the future. Networks experience growth, both in bandwidth at any particular point and in breadth—the number of devices connected and the physical (and/or logical—for instance, business units) area the network covers. Organizations must consider the following:

- How long can the organization expect to keep the capture solution?
- What are the bandwidth projections—for instance, when does the organization next plan to upgrade the network bandwidth?
- How much growth can occur within the current infrastructure—for instance, the organization uses what percentage of the current bandwidth?
- What network usage policy changes could influence the amount and types of traffic traversing the network—for instance, what if Facebook or YouTube is not allowed now, but is allowed in the future?
- What are the growth trends in the organization's network traffic and protocol-specific traffic, if available, or general trends outside the organization?

5 Real-World Examples

In this section, we present several examples to show how to apply the model to actual networks. In the first subsection, we discuss how we measured network traffic. In the last subsections, we discuss the examples. Each example presents the measurements from periodic observations over a 10- to 11-month span. We use these measurements to discuss storage requirements in a variety of scenarios—storing network flow only for all traffic, storing full packet capture for all traffic, and using a combination. For the first example, we also walk through the process of filling out the criticality and effectiveness ranking charts. Each example concludes with a discussion of how the organization would plan for future storage requirements based on the observed traffic trends and future network changes.

We chose the examples to show a variety of network sizes, based on the number of active IP addresses. The number of active IP addresses does not always correlate with how much traffic traverses a network's sensors. It is important to be aware that sensor placement influences how much traffic is seen. For instance, sensors inside a firewall see less traffic than sensors outside a firewall. Changing sensor placement within an organization requires recalculation of storage requirements.

5.1 Service and Attack Categories

Identifying appropriate traffic service and attack categories is important to the criticality and effectiveness ranking process. We categorized traffic services so that each category contained traffic that

- was useful for the same purposes, as identified in the *Network Data Usage Chart*
- could be used to compromise assets in the same way (e.g., all traffic in the category could be used to exfiltrate data and deliver malware)
- could be captured and categorized without the use of deep packet inspection

The 17 resulting categories of traffic were based first on protocol and then, if the protocol was TCP or UDP, broken down further into type of service. The protocols selected are

- TCP
- UDP
- Internet Control Message Protocol (ICMP)
- Encapsulating Security Payload (ESP)
- IPv4 (encapsulation only)
- IPv6
- Other

The category "Other" contains traffic with all other protocols. TCP and UDP break down further:

- TCP
  - HTTP: regular web traffic
  - Encrypted HTTP: encrypted web traffic such as HTTPS

  - Remote Connections: remote access protocols such as secure shell
  - Encrypted Tunneling: virtual private network connections
  - Email: email sending and retrieval
  - File Copy: remote file access protocols such as FTP
  - Encrypted File Copy: encrypted remote file access protocols such as FTPS
  - VoIP Signaling: voice over internet protocol administration traffic
  - Encrypted VoIP: encrypted voice over internet protocol administration traffic
  - Generic TCP: for all other TCP-related traffic
- UDP
  - DNS: domain name server protocol traffic
  - NTP: network time protocol traffic
  - Remote Connections: remote access protocols such as secure shell
  - Encrypted Tunneling: virtual private network connections
  - VoIP Signaling: voice over internet protocol administration traffic
  - Encrypted VoIP: encrypted voice over internet protocol administration traffic
  - Generic UDP: for all other UDP-related traffic

We selected the categories of attacks based on the attack chain [Hutchins 2011] and similarity of traffic during execution. The categories are

- Reconnaissance
  - Scanning
  - Indexing
- Delivery/Exploitation/Installation
  - Malware drop
  - Command and Control/backdoor communications
  - Worm propagation
- Command and Control
  - System control
- Action on Objectives
  - Exfiltration
  - Data corruption
  - DoS/DDoS flooding
  - DoS/DDoS crashing

5.2 Measurement Method

We used the SiLK tool set to measure network traffic. At each site, we obtained 10-11 months of data and divided it into service categories, as defined in the previous section.

To separate the traffic into the correct categories, we first looked at the protocol; if it was TCP or UDP, we examined the ports in use for each flow. For each TCP and UDP flow, we compared the source port and destination port and selected the numerically smaller as the most likely to indicate the service. We then compared the chosen port number to the ports that are officially associated

with the type of service. The ports used in the following examples for the above service categories are

- TCP
  - HTTP {80, 8080}
  - Encrypted HTTP {443}
  - Remote Connections {22, 23, 514, 3389, 5800}
  - Encrypted Tunneling {47, 1194, 1723}
  - Email {24, 25, 57, 109, 110, 143, 158, 209}
  - File Copy {20, 21, 115, 152, 427, 548}
  - Encrypted File Copy {992}
  - VoIP Signaling {1719, 1720, 2000, 2427, 2727, 5060}
  - Encrypted VoIP Signaling {2443, 5061}
  - Generic TCP for all other ports
- UDP
  - DNS {53}
  - NTP {123}
  - VoIP Signaling {1719, 2427, 2727, 5060}
  - Encrypted VoIP Signaling {2443, 5061}
  - Generic UDP for all other ports

This method, while simple to implement and quick to execute, does have some drawbacks. First, using port numbers to determine a service is not completely accurate. It is possible for any service to run on any port, though flows will most likely use their officially assigned port because that makes it most likely others can use the service. Second, this method does not allow for the categorization of traffic that occurs on ports that do not have standard assignment to a service. Third, it is possible that the service selected was not the lower port number of the source or destination ports. It would be more accurate, and much more complicated, to determine which flows were initiators of a session and use the destination ports of those flows to determine the service.

The goal for measuring traffic was to roughly estimate the types and amounts of traffic on a network to estimate the reduction in network capture data if different types of traffic were filtered out. The simplicity of our port selection method and the level of network traffic capture we could access (network flow with minimal metadata) met these requirements. Organizations whose network capture solution can determine the service or that can get the service information from the server and host logs can obtain greater precision.

Sizes for network and augmented flow records were calculated using NetFlow version 9 fields. The network flow size assumes these minimum fields: source address, destination address, source port, destination port, start time, end time, protocol, bytes, and packets. These fields make a record of 33 bytes [Powers 2010]. The augmented flow size assumes the minimum fields plus

- Flow direction: 1 byte
- Source AS: 2 bytes
- Destination AS: 2 bytes
- TCP flags: 1 byte

- TCP window size: 2 bytes
- ICMP type: 1 byte
- IP TTL minimum: 1 byte
- IP TTL maximum: 1 byte
- IP header length: 1 byte
- IP fragmentation flags: 1 byte
- IP total length minimum: 2 bytes
- IP total length maximum: 2 bytes

These fields result in an augmented flow record of 50 bytes.

We calculated network traffic trends per service category for all traffic and peak traffic. For each example, there were measurements for 10-11 months of traffic, collected in bytes per day per category. We imported the results into Microsoft Excel, putting each service in its own sheet and replacing dates with ordinal number of days. For instance, when the date range was March 1 through December 31, March 1 became day 1 and December 31 became day 306. Then we converted the byte values into megabytes to create a more manageable scale. Using the method explained in the Microsoft Office article "Modeling Exponential Growth" [Winston 2004], we used the ordinal days associated with traffic volumes to calculate an exponential growth formula for each category over the given time frame. Finally, we applied the compound growth rate formula

$$\left(\frac{\text{ending value}}{\text{starting value}}\right)^{\left(\frac{1}{\#\text{ months}}\right)} - 1$$

where the ending value is the value the exponential growth formula predicts for the last day (306 in the March through December instance), starting value is the value the exponential growth formula predicts for day 1, and # months is the length of the time range (10 in the example)

For the calculations of general traffic growth, we used each day's actual volume of traffic as the volume measurement. For the calculations of peak traffic growth, we followed the same process, except instead of using the original traffic volumes, each day's volume was replaced with the greatest volume seen until that date. So for instance, if the original volumes for days one through five were 2 GB, 1.5 GB, 3 GB, 1.5 GB, and 2.75 GB, they would become 2 GB, 2 GB, 3 GB, 3 GB, and 3 GB. This method of trending peaks may result in an overestimate of peak growth if the first few values in the series are much less than what would have been the peak value if the measurements had gone back further in time.

For example, Figure 3 and Figure 4 show the results of plotting one service for a 10-month time frame on a network—total daily volumes and peak volumes, respectively. Total daily volumes will seldom exhibit tight clustering around a trend line, but the trend line is good for averaging out traffic volumes to get an estimate of how they grow. The ordinal number of days, where day 1 is March 1 and day 306 is December 31, is the horizontal axis. The byte volume for a single service is the vertical axis. In Figure 3, even though the measurements vary greatly, there is a slight upward trend. Because Figure 4 considers only peak values seen in the traffic, it has the effect of smoothing the values and producing a scatterplot with a trend line that has a better fit, making the

upward trend more noticeable. In Figure 4, the first several values are much smaller than the later ones. If we had done measurements for one month before, the peak values coming into the first day of the 10 months shown would have been closer to 2.5.

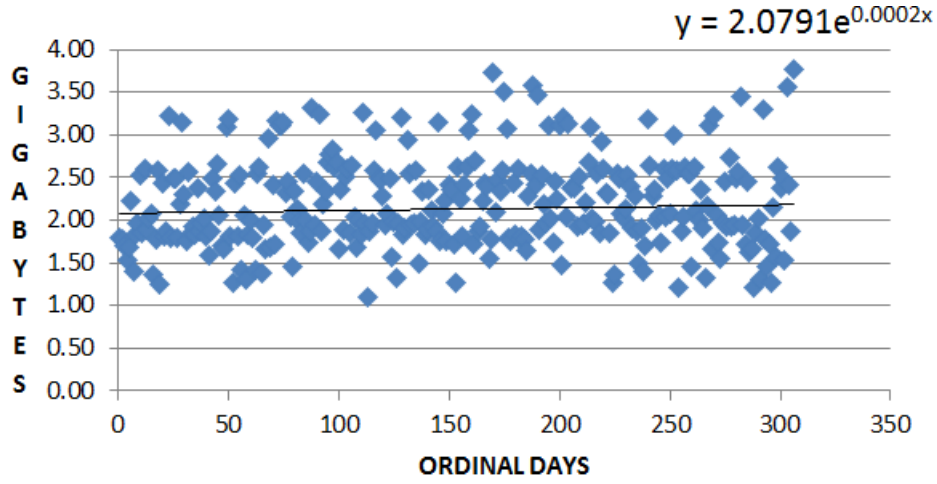

Figure 3: Example Total Daily Volume

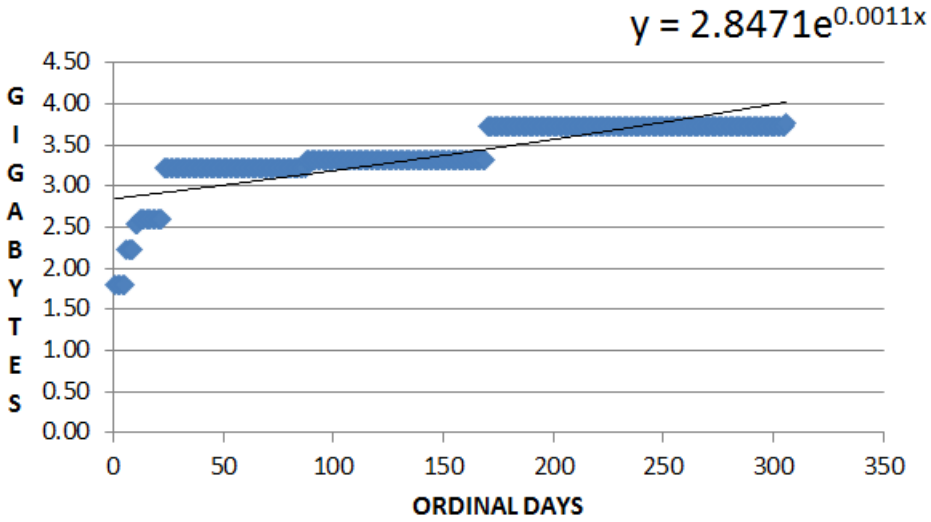

Figure 4: Example Peak Value Plot for Total Volumes Shown in Figure 3

We used the same process and formula to calculate the general and peak growth rates for the number of flows.

5.3 Small Private Network

This example is for a small private network. We analyzed traffic for this network from February through December 2013. During this time frame, there were several thousand active addresses and tens of terabytes of traffic a day that passed the border traffic capture sensors. The proportions of the traffic category volumes to total volume remained steady across all 10 months.

Table 12 shows the traffic values.

Table 12: Small Private Network Traffic Measurements

Service	Max Bytes (in MB)	Max Flows (1,000s)	Monthly Growth Rate %: All Traffic Bytes	Monthly Growth Rate %: Peak Traffic Bytes	Monthly Growth Rate %: All Traffic Flows	Monthly Growth Rate %: Peak Traffic Flows
TCP: HTTP	193660060.13	7889681.10	0.92%	2.47%	0.92%	2.47%
TCP: Encrypted HTTP	110162355.27	4210449.79	0.92%	2.47%	0.92%	2.47%
TCP: Remote Connections	47705167.47	303414.86	1.22%	2.78%	1.22%	2.47%
TCP: Encrypted Tunneling	8352.66	2051.57	2.47%	0.92%	0.92%	0.92%
TCP: Email	1913296.81	209125.57	0.92%	2.47%	0.92%	2.47%
TCP: File Copy	106127.67	36189.48	0.91%	2.78%	0.92%	2.47%
TCP: Encrypted File Copy	1.72	22.25	0.61%	1.85%	0.61%	1.85%
TCP: VoIP Signaling	3863.12	787.57	0.61%	3.41%	0.61%	2.16%
TCP: Encrypted VoIP Signaling	904.57	783.16	0.92%	2.47%	0.92%	2.47%
TCP: Other	85260447.29	1024270.04	0.92%	2.78%	0.92%	2.47%
UDP: DNS	936729.12	7425069.31	0.92%	2.78%	0.92%	2.78%
UDP: NTP	381739.18	24214.37	0.92%	2.78%	0.92%	2.47%
UDP: Remote Connections	3.93	25.85	0.92%	2.47%	0.92%	2.47%
UDP: Encrypted Tunneling	258997.37	274.47	1.22%	2.78%	0.92%	2.47%
UDP: VoIP Signaling	37343.89	78821.80	0.92%	2.47%	0.92%	2.47%
UDP: Encrypted VoIP Signaling	384.27	924.88	0.61%	1.85%	0.61%	1.85%
UDP: Other	83703584.19	791745.10	0.92%	2.47%	0.61%	2.16%
ICMP	47631.54	215569.00	1.23%	2.78%	0.92%	2.47%
ESP	2862233.09	18697.43	0.92%	2.78%	0.92%	2.78%
IPv4	0.08	1.94	0.92%	2.47%	0.92%	2.16%
IPv6	176.22	323.48	0.92%	2.47%	0.92%	2.47%
Other	2.08	4.61	0.31%	1.54%	0.61%	1.85%

5.3.1 Step 1: Network Data Usage

This organization is interested in investigating attacks and understanding their network, so their *Network Data Usage* chart looks like Table 13.

Table 13: Small Organization Example Network Data Usage

Y/N	Reason Number	Purpose	Capture Method*	Who	How Much	When	How Long	Using What	Transfer-ring What	How
	1	Investigate attacks	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓	✓	✓	
Y			P	✓	✓	✓	✓	✓	✓	✓
	2	Police policies	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓	✓	✓	
__			P	✓	✓	✓	✓	✓	✓	✓
	3	Provide information to create policies	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓	✓	✓	
__			P	✓	✓	✓	✓	✓	✓	✓
	4	Understand normal network traffic	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓			
Y			P	✓	✓	✓	✓			
	5	Understand abnormal network traffic	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓			
Y			P	✓	✓	✓	✓			
	6	Plan for network upgrades	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓			
Y			P	✓	✓	✓	✓			

*** N = network flow, A = augmented flow, P = pcap**

5.3.2 Step 2: Number of Storage Days

The organization would like to be able to look back at traffic for at least a half a year when analyzing traffic for situational awareness purposes, so its *Number of Storage Days* chart looks like Table 14.

Table 14: Small Organization Example Number of Storage Days

Capture Method	Selected Network Data Usage Purpose	Days of Traffic Needed	Days to Store
Pcap	1	To investigate attacks: ___	See Attack/Risk and Requirements Charts
	2	To police policies based on traffic content: ___	Consider using filtering/IDS
	3	To see what content is passing to create network use policies: ___	Consider using filtering/IDS
Augmented Flow	1	To investigate attacks: ___	See Attack/Risk and Requirements Charts
	2	To police policies based on traffic content: ___	Consider using filtering/IDS
	3	To see what content is passing to create network use policies: ___	Consider using filtering/IDS
Network Flow	1	To investigate attacks: ___	See Attack/Risk and Requirements Charts
	2	To see volume, who, and when information to police policies based on traffic content after the fact: ___	____
	3	To see volume, who, and when information to create network use policies: ___	____
	4, 5	To trend traffic to understand normal and abnormal traffic: ___	180
	6	To trend traffic to plan for upgrades: ___	180

5.3.3 Step 3: Attack Type Criticality

The *Attack Type Criticality Chart* for this organization looks like Table 15.

Table 15: Small Organization Example Attack Type Criticality Chart

		Criticality of determining:		
		Network Flow (N)	**Augmented Flow (A)**	**Pcap (P)**
		who is affected, how much, when, how often	**using what, transferring what**	**how and what data**
Attack Category	**Scanning**	2	1	1
	Indexing	2	1	1
	Malware Drop	3	3	3
	C2/Backdoor Communications	3	3	2
	Worm Propagation	3	3	3
	System Control	3	3	2
	Exfiltration	3	3	2
	Data Corruption	2	2	2
	DoS/DDoS Flooding	3	2	1
	DoS/DDoS Crashing	3	2	1

Criticality Levels
3: Legal ramifications/major increase in financial losses
2: Increase in financial losses or time to resolve issues caused by the incident
1: None or minor inconvenience

5.3.4 Step 4: Effectiveness

For evaluating effectiveness, the organization defines time ranges as one day, up to one week, up to four weeks, up to 12 weeks, and up to 24 weeks. The organization's *Effectiveness Chart* looks like Table 16.

Table 16: Small Organization Example Effectiveness Chart

	Column Value:	1 day	7 days	28 days	84 days	168 days	
	Effectiveness Time Range (in days):	**Immediate**	**Short-term**	**Mid-term**	**Long-term**	**Archive**	**Days of Greatest Effectiveness**
Attack Category	**Scanning**	3	3	1	1	1	7
	Indexing	3	3	1	1	1	7
	Malware Drop	1	4	5	4	4	28
	C2/Backdoor Communications	3	5	4	3	3	7
	Worm Propagation	4	5	4	1	1	7
	System Control	3	4	5	5	5	168
	Exfiltration	3	4	5	5	5	168
	Data Corruption	5	5	5	5	5	168
	DoS/DDoS Flooding	4	3	1	1	1	1
	DoS/DDoS Crashing	4	3	1	1	1	1

Effectiveness Value
5: Five times the effectiveness of value 1—data at this time frame is critical
4: Four times the effectiveness of value 1
3: Three times the effectiveness of value 1
2: Two times the effectiveness of value 1
1: Data at this time frame has little use

5.3.5 Step 5 and 6: Essentiality and Storage Requirements

Using the values from the previous charts and doing the calculations as described in Section 4.5 results in the *Risk and Requirements Chart* shown in Table 17. For the full packet capture storage per day calculation, file size was assumed to be 50 MB, and the file overhead for Wireshark's .pcap format was used for the formula P = ((Bytes / File Size) x Overhead) + Bytes. The organization wants to store all traffic as network flow; all unencrypted traffic, email, or ESP as augmented flow; and all traffic that is not encrypted, email, ESP, or VoIP as full packet capture. The organization's savings in augmented flow storage is about 20%, and its savings for full packet capture storage is about 22%. The crossed-out numbers in the "Storage per Day" columns indicate that this organization will not store that service at the corresponding tier. The crossed-out numbers in the "Totals" row represent the value of storing all services at the tier, while the other numbers represent the value of storing only the selected services.

Table 17: Small Organization Example Risk and Requirements Chart

Traffic Type/Service	Attack Category/Storage Tier: Scanning N	Scanning A	Scanning P	Indexing N	Indexing A	Indexing P	Malware Drop N	Malware Drop A	Malware Drop P	C2/Backdoor Communications N	C2/Backdoor Communications A	C2/Backdoor Communications P	Worm Propagation N	Worm Propagation A	Worm Propagation P	System Control N	System Control A	System Control P	Exfiltration N	Exfiltration A	Exfiltration P	Data Corruption N	Data Corruption A	Data Corruption P	DoS/DDoS Flooding N	DoS/DDoS Flooding A	DoS/DDoS Flooding P	DoS/DDoS Crashing N	DoS/DDoS Crashing A	DoS/DDoS Crashing P	Risk Value N	Risk Value A	Risk Value P	Essentiality N	Essentiality A	Essentiality P	Effectiveness	Max Expected Traffic per Day: Useful days	Bytes	Flows	Storage per Day: N (GB)	A (GB)	P (TB)
TCP: HTTP	2	1	1	2	1	1	3	3	3	3	3	2	3	3	3	3	3	2	3	3	2	2	2	2	3	2	1	3	2	1	**1**	1	1	2	2	2	168	220601322365874	8528747774	262.2	397.2	200.6
TCP: Encrypted HTTP	2			2			3			3			3			3			3						3			3			**1**	1	1	2	-1	-1	168	126223363370437	4489225472	138.0	~~209.0~~	~~114.8~~
TCP: Remote Connections	2	1	1	2	1	1	3	3	3	3	3	2	3	3	3	3	3	2	3	3	2	2	2	2	3	2	1	3	2	1	**1**	1	**1**	2	2	2	168	50323857314568	340708532	10.5	15.9	45.8
TCP: Encrypted Tunneling	2			2			3			3			3			3			3						3			3			**1**	**1**	1	2	-1	-1	168	8772295742	2149679	0.1	~~0.1~~	~~0.0~~
TCP: Email	2	1	1	2	1	1	3	3	3	3	3	2	3	3	3	3	3	2	3	3	2	2	2	2	3	2	1	3	2	1	**1**	3	3	2	0	0	168	2179235969659	238000816	7.3	~~11.1~~	~~2.0~~
TCP: File Copy							3	3	3	3	3	2	3	3	3				3	3	2	2	2	2							**1**	2	2	2	1	1	168	113930303698	41390251	1.3	1.9	0.1
TCP: Encrypted File Copy							3			3			3						3												**1**	1	1	2	-1	-1	168	1845132	23137	0.0	~~0.0~~	~~0.0~~
TCP: VoIP	2	1	1	2	1	1	3	3	3	3	3	2	3	3	3	3	3	2	3	3	2	2	2	2	3	2	1	3	2	1	**1**	1	3	2	2	0	168	4149827048	902004	0.0	0.0	~~0.0~~
TCP: Encrypted VoIP	2	1		2	1		3	3		3	3		3	3		3	3		3	3		2	2		3	2		3	2		**1**	1	3	2	2	-3	168	953552148	785247	0.0	~~0.0~~	~~0.0~~
TCP: Generic (all other services)	2	1	1	2	1	1	3	3	3	3	3	2	3	3	3	3	3	2	3	3	2	2	2	2	3	2	1	3	2	1	**1**	1	1	2	2	2	168	92348254561301	1109477172	34.1	51.7	84.0
UDP: DNS				2	1	1	3	3	3	3	3	2	3	3	3	3	3	2	3	3	2	2	2	2	3	2	1	3	2	1	**1**	1	1	2	2	2	168	1007875118052	7651023301	235.1	356.3	0.9
UDP: NTP	2	1	1							3	3	2				3	3	2							3	2	1	3	2	1	**1**	1	1	2	2	1	168	400462075300	26338476	0.8	1.2	0.4
UDP: Remote Connections	2	1	1	2	1	1	3	3	3	3	3	2	3	3	3	3	3	2	3	3	2	2	2	2	3	2	1	3	2	1	**1**	1	1	2	2	2	168	4772542	28493	0.0	0.0	0.0
UDP: Encrypted Tunneling	2			2			3			3			3			3			3						3			3			**1**	1	1	2	-1	-1	168	284631376818	300709	0.0	~~0.0~~	~~0.3~~
UDP: VoIP	2	1	1	2	1	1	3	3	3	3	3	2	3	3	3	3	3	2	3	3	2	2	2	2	3	2	1	3	2	1	**1**	1	3	2	2	0	168	41297612750	83656842	2.6	3.9	~~0.0~~
UDP: Encrypted VoIP	2	1		2	1		3	3		3	3		3	3		3	3		3	3		2	2		3	2		3	2		**1**	1	3	2	2	-3	168	444285998	1019663	0.0	~~0.0~~	~~0.0~~
UDP: Generic (all other services)	2	1	1	2	1	1	3	3	3	3	3	2	3	3	3	3	3	2	3	3	2	2	2	2	3	2	1	3	2	1	**1**	1	1	2	2	2	168	95409111138888	829502665	25.5	38.6	86.8
ICMP	2	1		2	1		3	3	3	3	3	2	3	3	3	3	3	2	3	3	2	2	2	2	3	2	1	3	2	1	**1**	**1**	**1**	2	2	2	168	51760537767	228443218	7.0	10.6	0.0
ESP	2			2			3			3			3			3			3						3			3			**1**	1	1	2	-1	-1	168	3296418615180	20503897	0.6	~~1.0~~	~~3.0~~
IPv4	2	1	1	2	1	1	3	3	3	3	3	2	3	3	3	3	3	2	3	3	2	2	2	2	3	2	1	3	2	1	**1**	1	1	2	2	2	168	87260	2052	0.0	0.0	0.0
IPv6	2	1	1	2	1	1	3	3	3	3	3	2	3	3	3	3	3	2	3	3	2	2	2	2	3	2	1	3	2	1	**1**	1	1	2	2	2	168	201631194	352194	0.0	0.0	0.0
Other	2	1	1	2	1	1	3	3	3	3	3	2	3	3	3	3	3	2	3	3	2	2	2	2	3	2	1	3	2	1	**1**	1	1	2	2	2	168	2249085	4752	0.0	0.0	0.0
																																						Totals:		725.1	~~1098.7~~ 877.5	~~538.7~~ 418.6
																																						Storage Days Goal:		180	168	180
																																						Savings:			20.13%	22.29%

N = Network Flow, A = Augmented Flow, P = Pcap

Risk Values

5: Storage of information has legal ramifications that cannot be mitigated.

4: Storage or leakage of information has legal or major financial ramifications that can be mitigated, but mitigation is not already in place.

3: Storage or leakage of information has financial ramifications that can be mitigated, but mitigation is not already in place.

2: Storage or leakage of information has legal or financial ramifications, but mitigations are in place.

1: Storage or leakage of information would cause minor inconvenience. This value is used anywhere the ramifications are unlikely to be anything other than leaked network structure if the stored data was subject to unauthorized access.

To project required storage, it is possible to use either the growth rate from all traffic or that calculated from the change in peaks. Using the growth rate of peak traffic better ensures that the actual storage days will always meet the desired storage period. If the storage days goal is a target and not a hard number, organizations may find the total volume growth rate acceptable—they would usually meet the storage time goal, but occasionally storage time would decrease if peaks occurred several days during the period. Looking out 24 months, this organization's projections would result in the data in Table 18. The table shows values for both peak and total volume calculations to illustrate the differences between the two methods. The values that are crossed-out in the service rows represent the values that this organization will not store. The crossed-out numbers in the "Totals" row represent the value of storing all services at the tier, while the other numbers represent the value of storing only the selected services.

Table 18: Small Organization Storage Projections

	Monthly Growth Rates				Storage per Day in 24 Months: All			Storage per Day in 24 Months: Peak		
	All Traffic (Bytes)	**Peak Traffic (Bytes)**	**All Traffic (Flows)**	**Peak Traffic (Flows)**	**N (GB)**	**A (GB)**	**P (TB)**	**N (GB)**	**A (GB)**	**P (TB)**
TCP: HTTP	0.92%	2.47%	0.92%	2.47%	292.7	443.3	223.9	351.4	532.3	268.9
TCP: Encrypted HTTP	0.92%	2.47%	0.92%	2.47%	154.0	~~233.3~~	~~128.1~~	184.9	~~280.2~~	~~153.9~~
TCP: Remote Connections	1.22%	2.78%	1.22%	2.47%	12.1	18.4	52.9	14.0	21.3	63.6
TCP: Encrypted Tunneling	2.47%	0.92%	0.92%	0.92%	0.1	~~0.1~~	~~< 0.1~~	0.1	~~0.1~~	~~< 0.1~~
TCP: Email	0.92%	2.47%	0.92%	2.47%	8.2	~~12.4~~	~~2.2~~	9.8	~~14.9~~	~~2.7~~
TCP: File Copy	0.91%	2.78%	0.92%	2.47%	1.4	2.2	0.1	1.7	2.6	0.1
TCP: Encrypted File Copy	0.61%	1.85%	0.61%	1.85%	< 0.1	~~< 0.1~~	~~< 0.1~~	< 0.1	~~< 0.1~~	~~< 0.1~~
TCP: VoIP	0.61%	3.41%	0.61%	2.16%	< 0.1	< 0.1	~~< 0.1~~	< 0.1	0.1	~~< 0.1~~
TCP: Encrypted VoIP	0.92%	2.47%	0.92%	2.47%	< 0.1	~~< 0.1~~	~~< 0.1~~	< 0.1	~~< 0.1~~	~~< 0.1~~
TCP: Generic (all other services)	0.92%	2.78%	0.92%	2.47%	38.1	57.7	93.7	45.7	69.2	116.7
UDP: DNS	0.92%	2.78%	0.92%	2.78%	262.5	397.7	1.0	326.8	495.1	1.3
UDP: NTP	0.92%	2.78%	0.92%	2.47%	0.9	1.4	0.4	1.1	1.6	0.5
UDP: Remote Connections	0.92%	2.47%	0.92%	2.47%	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
UDP: Encrypted Tunneling	1.22%	2.78%	0.92%	2.47%	< 0.1	~~< 0.1~~	~~0.3~~	< 0.1	~~0.1~~	~~0.4~~
UDP: VoIP	0.92%	2.47%	0.92%	2.47%	2.9	4.3	~~< 0.1~~	3.4	5.2	~~0.1~~
UDP: Encrypted VoIP	0.61%	1.85%	0.61%	1.85%	< 0.1	~~0.1~~	~~< 0.1~~	< 0.1	~~0.1~~	~~< 0.1~~
UDP: Generic (all other services)	0.92%	2.47%	0.61%	2.16%	27.4	41.6	96.9	32.9	49.9	116.3
ICMP	1.23%	2.78%	0.92%	2.47%	7.8	11.9	0.1	9.4	14.3	0.1
ESP	0.92%	2.78%	0.92%	2.78%	0.7	~~1.1~~	~~3.3~~	0.9	~~1.3~~	~~4.2~~
IPv4	0.92%	2.47%	0.92%	2.16%	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
IPv6	0.92%	2.47%	0.92%	2.47%	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
Other	0.31%	1.54%	0.61%	1.85%	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
				Totals	808.8	~~1225.3~~ 978.3	~~603.1~~ 469.1	982.3	~~1494.0~~ 1197.4	~~728.6~~ 567.5

Using the "Storage Days Goal" values for each tier would result in the total storage needs shown in Table 19. The daily storage needs for the "Requirements with Current Values" column and all

the number of days values come from the corresponding values in Table 17. The daily storage needs for the last two columns come from the corresponding values in Table 18.

Table 19: Small Organization Total Storage Requirements

Storage Tier	Requirements with Current Values	Requirements with All Traffic Projected Values	Requirements with Peak Traffic Projected Values
Network Flow	180 days X 725.1 = 130,518 GB	180 days X 808.8 = 145,584 GB	180 days X 986.2 = 177,516 GB
Augmented Flow	168 days X 887.5 = 149,100 GB	168 days X 978.3 = 164,354.4 GB	168 days X 1197.4 = 201,163.2 GB
Pcap	180 days X 418.6 = 75,348 GB	180 days X 469.1 = 84,438 TB	180 days X 567.5 = 102,150 TB

5.4 Medium Private Network

This example is for a medium-sized private network. We analyzed traffic for this network from June 2013 through March 2014. During the measurement time frame, there were more than 222,000 active addresses of traffic per day that passed the border traffic capture sensors. Table 20 shows traffic values.

Table 20: Medium Private Network Traffic Measurements

Service	Max Bytes (in MB)	Max Flows (1,000s)	Monthly Growth Rate %: All Traffic Bytes	Monthly Growth Rate %: Peak Traffic Bytes	Monthly Growth Rate %: All Traffic Flows	Monthly Growth Rate %: Peak Traffic Flows
TCP: HTTP	26038567.78	653884.8	-5.88%	13.22%	-11.41%	0.91%
TCP: Encrypted HTTP	1898350.03	35889.78	-5.88%	2.45%	-0.60%	2.45%
TCP: Remote Connections	3066185.66	11798.62	-5.88%	9.18%	8.52%	11.86%
TCP: Encrypted Tunneling	1374.22	784.45	-8.69%	18.49	22.88%	24.00%
TCP: Email	632840.76	4642.04	24.38%	63.87%	0.91%	19.57%
TCP: File Copy	2378301.12	79234.34	-14.06%	14.26%	-16.62%	4.97%
TCP: Encrypted File Copy	1.67	32.03	No Growth	< 0.01%	No Growth	0.24%
TCP: VoIP Signaling	3251.44	850.65	-8.69%	4.97%	0.61%	22.14%
TCP: Encrypted VoIP Signaling	2786.19	33.87	-16.62%	23.63%	0.02%	< 0.01%
TCP: Other	18171514.76	139301.49	-2.98%	9.18%	-5.88%	< 0.01%
UDP: DNS	576607.26	311526.85	-2.98%	13.23%	-5.88%	2.14%
UDP: NTP	170582.67	1156.42	5.61%	110.09%	1.83%	14.96%
UDP: Remote Connections	263.89	38.62	-8.69%	78.92%	-8.69%	58.02%
UDP: Encrypted Tunneling	5883.84	1.86	19.57%	97.14%	-0.02%	21.03%
UDP: VoIP Signaling	837.48	1522.67	0.61%	12.89%	0.91%	10.18%
UDP: Encrypted VoIP Signaling	249.97	149.38	4.33%	35.56%	4.97%	21.77%
UDP: Other	1512119.81	7085.30	-11.41%	12.20%	-0.24%	7.22%
ICMP	2143.02	4019.67	-0.27%	7.22%	-11.41%	4.02%
ESP	1630035.60	2851.99	3.39%	7.54%	-45.45%	4.65%
IPv4	0.01	0.03	1.53%	2.45%	No Growth	1.83%
IPv6	159.12	9.32	-21.53%	-14.06%	53.30%	12.20%
Other	11007.43	439.22	No Growth	No Growth	46.93%	1.81%

Table 21 maps the storage requirements at each tier for several filtering options. For this organization's resources and needs, full packet capture is reasonable for storing all but encrypted data, email, DNS, NTP, ICMP, and VoIP; augmented flow for storing all but encrypted data; and network flow for storing all traffic. The organization needs to plan for about 50.46 TB of storage for each day's worth of data it needs to store.

Table 21: Medium Private Network Raw Data Network Storage Requirement Examples

Traffic Kept	Size for Network Flow (GB)	Size for Augmented Flow (GB)	Size for Pcap (GB)
All	38.58	58.45	54,788.15
All but Encrypted	37.44	56.73	52,924.24
All but Encrypted and Email	37.30	56.52	52,306.23
All but Encrypted, Email, DNS, NTP, and ICMP	27.57	41.77	51,574.46
All but Encrypted, Email, DNS, NTP, ICMP, and VoIP	27.50	41.66	51,570.47

We calculated size for network flow as

$$\text{Flows} \times \text{Network Flow Record Size}$$

where Flows come from the "Max Flows" column in Table 24 and Network Flow Record Size is 33 bytes

We calculated size for augmented flow as

$$\text{Flows} \times \text{Augmented Flow Record Size}$$

where Flows come from the "Max Flows" column in Table 24 and Augmented Flow Record Size is 50 bytes, as explained in Section 5.2

We calculated size for full packet capture as

$$((\text{Bytes} / \text{File Size}) \times \text{Overhead}) + \text{Bytes}$$

where Bytes comes from Table 24, File Size is how much raw data to store per file, and Overhead is the file size overhead for the file storage format calculated from Table 4

For this example, we used 50 MB as the file size and .pcapng as the file format, meaning there is a 759 KB overhead. Table 22 shows the savings for the different scenarios.

Table 22: Medium Private Network Storage Savings Examples

Traffic Kept	Savings for Network Flow	Savings for Augmented Flow	Savings for Pcap
All	-	-	-
All but Encrypted	2.95%	2.94%	3.40%
All but Encrypted and Email	3.32%	3.30%	3.40%
All but Encrypted, Email, DNS, NTP, and ICMP	28.54%	28.54%	5.87%
All but Encrypted, Email, DNS, NTP, ICMP, and VoIP	28.72%	28.73%	5.87%

Using the growth rates for this network, Table 23 shows the projected storage required for one day's worth of data in two years. We made the calculations for all traffic changes, as well as peak traffic changes, as explained in Step 6d. This example is interesting because many of the services have decreased use, while others grow rapidly. Even when storage needs will decrease in the future, it is still important to have at least the storage necessary for current needs. This example also dramatically illustrates why it is important to consider peak traffic growth. Even the traffic that decreases in general increases in peak volumes.

Table 23: Medium Private Network Two-Year Growth Projections

	Monthly Growth Rates				Storage per Day in 24 Months: All			Storage per Day in 24 Months: Peak		
	All Traffic (Bytes)	**Peak Traffic (Bytes)**	**All Traffic (Flows)**	**Peak Traffic (Flows)**	**N (MB)**	**A (MB)**	**P (GB)**	**N (MB)**	**A (MB)**	**P (TB)**
TCP: HTTP	-5.88%	13.22%	-11.41%	0.91%	1123.69	1702.57	5938.59	25576.21	38751.84	488.85
TCP: Encrypted HTTP	-5.88%	2.45%	-0.60%	2.45%	977.59	1481.20	432.95	2019.17	3059.34	3.24
TCP: Remote Connections	-5.88%	9.18%	8.52%	11.86%	2642.29	4003.47	699.30	5469.41	8286.99	24.07
TCP: Encrypted Tunneling	-8.69%	18.49%	22.88%	24.00%	3467.59	5253.93	0.15	4311.22	6532.14	0.08
TCP: Email	24.38%	63.87%	0.91%	19.57%	181.57	275.11	116146.94	10655.07	16144.05	84855.78
TCP: File Copy	-14.06%	14.26%	-16.62%	4.97%	31.79	48.17	61.19	7987.16	12101.75	55.61
TCP: Encrypted File Copy	No growth	< 0.01%	No growth	0.24%	1.01	1.53	0.00	1.07	1.62	0.00
TCP: VoIP	-8.69%	4.97%	0.61%	22.14%	30.98	46.94	0.36	3252.78	4928.46	0.01
TCP: Encrypted VoIP	-16.62%	23.63%	0.02%	< 0.01%	1.07	1.62	0.03	1.07	1.62	0.43
TCP: Generic (all other services)	-2.98%	9.18%	-5.88%	< 0.01%	1023.85	1551.29	8585.42	4383.99	6642.41	142.63
UDP: DNS	-2.98%	13.23%	-5.88%	2.14%	2289.68	3469.22	272.43	16297.09	24692.56	10.85
UDP: NTP	5.61%	110.09%	1.83%	14.96%	56.24	85.21	617.38	1033.12	1565.34	8893324.69
UDP: Remote Connections	-8.69%	78.92%	-8.69%	58.02%	0.14	0.21	0.03	71419.31	108211.08	291.50
UDP: Encrypted Tunneling	19.57%	97.14%	-0.02%	21.03%	0.06	0.09	419.08	5.71	8.66	66627.73
UDP: VoIP	0.61%	12.89%	0.91%	10.18%	59.56	90.24	0.95	490.89	743.78	0.01
UDP: Encrypted VoIP	4.33%	35.56%	4.97%	21.77%	15.06	22.82	0.68	531.10	804.69	0.35
UDP: Generic (all other services)	-11.41%	12.20%	-0.24%	7.22%	210.49	318.92	80.63	1188.21	1800.31	22.85
ICMP	-0.27%	7.22%	-11.41%	4.02%	6.91	10.47	1.96	325.77	493.59	0.01
ESP	3.39%	7.54%	-45.45%	4.65%	0.00	0.00	3543.09	267.18	404.82	8.90
IPv4	1.53%	2.45%	No growth	1.83%	0.00	0.00	0.00	0.00	0.00	0.00
IPv6	-21.53%	-14.06%	53.30%	12.20%	8324.18	12612.40	0.00	4.65	7.04	0.00
Other	No growth	No growth	46.93%	1.81%	141660.12	214636.55	0.01	21.26	32.21	0.01

5.5 Large Private Network

This example is for a large-sized private network. We analyzed traffic for this network from February through December 2013. During the measurement time frame, there were more than 642,800 active IP addresses and thousands of terabytes of traffic a day that passed the border traffic capture sensors. Table 24 shows the traffic values.

Table 24: Large Private Network Traffic Measurements

Service	Max Bytes (in MB)	Max Flows (1,000s)	Monthly Growth Rate %: All Traffic Bytes	Monthly Growth Rate %: Peak Traffic Bytes	Monthly Growth Rate %: All Traffic Flows	Monthly Growth Rate %: Peak Traffic Flows
TCP: HTTP	527910228.29	20021961.98	1.54%	0.61%	1.54%	0.61%
TCP: Encrypted HTTP	233048709.85	14256626.69	1.85%	0.92%	1.85%	0.92%
TCP: Remote Connections	2247755.58	227958.88	0.92%	0.92%	0.92%	0.61%
TCP: Encrypted Tunneling	167423.88	65622.15	1.85%	0.61%	2.47%	0.92%
TCP: Email	3186535.75	432854.20	1.23%	0.61%	1.85%	9.58%
TCP: File Copy	1000390.47	23550.00	1.85%	0.92%	1.85%	0.61%
TCP: Encrypted File Copy	0.53	6.25	No Growth	1.54%	1.23%	1.85%
TCP: VoIP Signaling	716.94	3399.42	1.85%	0.61%	1.85%	1.23%
TCP: Encrypted VoIP Signaling	2159.59	474.78	1.54%	0.92%	1.85%	0.61%
TCP: Other	12137593.43	2188152.27	1.85%	0.92%	1.54%	0.92%
UDP: DNS	984418.01	8547858.43	1.54%	0.92%	1.54%	0.92%
UDP: NTP	1056.74	10203.43	1.23%	0.92%	1.23%	0.92%
UDP: Remote Connections	5.65	22.90	No Growth	0.92%	1.23%	0.61%
UDP: Encrypted Tunneling	25.15	114.23	1.85%	0.61%	1.54%	0.92%
UDP: VoIP Signaling	115851.74	270620.34	1.54%	0.92%	1.54%	0.92%
UDP: Encrypted VoIP Signaling	1973.97	4753.30	0.15%	0.61%	0.15%	0.61%
UDP: Other	6560407.74	833223.09	1.54%	0.92%	1.23%	0.92%
ICMP	25618.16	102972.74	1.85%	0.61%	1.54%	0.92%
ESP	3702190.59	577.62	0.92%	0.61%	1.54%	0.92%
IPv4	0.64	13.88	0.92%	0.31%	0.61%	0.31%
IPv6	2601.17	0.46	No Growth	2.47%	No Growth	0.61%
Other	1443056.78	122.40	1.85%	0.31%	1.54%	0.92%

Table 25 maps the storage requirements at each tier for several filtering options. For this organization's resources and needs, full packet capture is reasonable for storing all but encrypted data, email, DNS, NTP, ICMP, and VoIP; augmented flow for storing all but encrypted data; and network flow for storing all traffic. The organization needs to plan for about 533 TB of storage for each day's worth of data it needs to store.

Table 25: Large Private Network Storage Requirement Examples

Traffic Kept	Size for Network Flow (TB)	Size for Augmented Flow (TB)	Size for Pcap (TB)
All	1.41	2.14	755.84
All but Encrypted	0.98	1.49	533.42
All but Encrypted and Email	0.97	1.47	530.38
All but Encrypted, Email, DNS, NTP, and ICMP	0.71	1.07	529.41

All but Encrypted, Email, DNS, NTP, ICMP, and VoIP	0.70	1.06	529.30

We calculated size for network flow as

$$\text{Flows} \times \text{Network Flow Record Size}$$

where Flows come from the "Max Flows" column in Table 24 and Network Flow Record Size is 33 bytes

We calculated size for augmented flow as

$$\text{Flows} \times \text{Augmented Flow Record Size}$$

where Flows come from the "Max Flows" column in Table 24 and Augmented Flow Record Size is 50 bytes, as explained in Section 5.2

We calculated size for full packet capture as

$$((\text{Bytes} / \text{File Size}) \times \text{Overhead}) + \text{Bytes}$$

where Bytes comes from Table 24, File Size is how much raw data to store per file, and Overhead is the file size overhead for the file storage format calculated from Table 4

For this example, we used 50 MB as the file size and .pcap as the file format, meaning there is a 1 KB overhead. Table 26 shows the savings for the different scenarios.

Table 26: Large Private Network Storage Savings Examples

Traffic Kept	Savings for Network Flow	Savings for Augmented Flow	Savings for Pcap
All	-	-	-
All but Encrypted	30.49%	30.49%	29.43%
All but Encrypted and Email	31.41%	31.41%	29.83%
All but Encrypted, Email, DNS, NTP, and ICMP	49.84%	49.84%	29.96%
All but Encrypted, Email, DNS, NTP, ICMP, and VoIP	50.43%	50.43%	29.97%

Using the growth rates for this network, Table 27 shows the projected storage required for one day's worth of data in two years. We made the calculations for all traffic changes and peak traffic changes, as explained in Step 6d.

Table 27: Large Private Network Two-Year Growth Projections

	Monthly Growth Rates				Storage per Day in 24 Months: All			Storage per Day in 24 Months: Peak		
	All Traffic (Bytes)	**Peak Traffic (Bytes)**	**All Traffic (Flows)**	**Peak Traffic (Flows)**	**N (GB)**	**A (GB)**	**P (TB)**	**N (GB)**	**A (GB)**	**P (TB)**
TCP: HTTP	1.54%	0.61%	1.54%	0.61%	739.2	1120.0	604.5	661.9	1002.9	541.4
TCP: Encrypted HTTP	1.85%	0.92%	1.85%	0.92%	546.0	827.2	276.8	489.1	741.0	248.0
TCP: Remote Connections	0.92%	0.92%	0.92%	0.61%	7.8	11.8	2.4	7.5	11.4	2.4
TCP: Encrypted Tunneling	1.85%	0.61%	2.47%	0.92%	2.7	4.1	0.2	2.3	3.4	0.2
TCP: Email	1.23%	0.61%	1.85%	9.58%	16.6	25.1	3.5	39.9	60.4	3.3
TCP: File Copy	1.85%	0.92%	1.85%	0.61%	0.9	1.4	1.2	0.8	1.2	1.1
TCP: Encrypted File Copy	0.00%	1.54%	1.23%	1.85%	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
TCP: VoIP	1.85%	0.61%	1.85%	1.23%	0.1	0.2	< 0.1	0.1	0.2	< 0.1
TCP: Encrypted VoIP	1.54%	0.92%	1.85%	0.61%	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
TCP: Generic (all other services)	1.85%	0.92%	1.54%	0.92%	80.8	122.4	14.4	75.1	113.7	12.9
UDP: DNS	1.54%	0.92%	1.54%	0.92%	315.6	478.2	1.1	293.2	444.3	1.0
UDP: NTP	1.23%	0.92%	1.23%	0.92%	0.4	0.6	< 0.1	0.4	0.5	< 0.1
UDP: Remote Connections	0.00%	0.92%	1.23%	0.61%	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
UDP: Encrypted Tunneling	1.85%	0.61%	1.54%	0.92%	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
UDP: VoIP	1.54%	0.92%	1.54%	0.92%	10.0	15.1	0.1	9.3	14.1	0.1
UDP: Encrypted VoIP	0.15%	0.61%	0.15%	0.61%	0.1	0.2	< 0.1	0.2	0.2	< 0.1
UDP: Generic (all other services)	1.54%	0.92%	1.23%	0.92%	29.7	44.9	7.5	28.6	43.3	7.0
ICMP	1.85%	0.61%	1.54%	0.92%	3.8	5.8	< 0.1	3.5	5.4	< 0.1
ESP	0.92%	0.61%	1.54%	0.92%	< 0.1	< 0.1	3.9	< 0.1	< 0.1	3.8
IPv4	0.92%	0.31%	0.61%	0.31%	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
IPv6	0.00%	2.47%	0.00%	0.61%	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
Other	1.85%	0.31%	1.54%	0.92%	< 0.1	< 0.1	1.7	< 0.1	< 0.1	1.4

6 Conclusion

Many organizations utilizing or needing to implement network traffic monitoring for security or general network awareness are finding that they cannot store full packet capture of traffic for a time span that permits effective incident response, forensic analysis, and other desired traffic analysis. With high-speed networks and ever-increasing network traffic volumes, this problem is getting worse. These organizations need a solution that provides them the information needed for a reasonable time frame without requiring exorbitant amounts of storage, for which the initial expenditure for hardware or the ongoing maintenance costs may be prohibitively expensive.

Traffic captured as network flow and, often, augmented flow requires much less storage and maintenance. This storage savings makes these options attractive to organizations looking for a solution other than full packet capture. Unfortunately these capture methods do not always provide the information needed to resolve or investigate an incident or to give a detailed view of what is going on in a network. Organizations need a network traffic capture solution that provides the benefits of each capture method.

One possible solution is to use all three methods selectively, storing only certain types of traffic at the more storage-intensive tiers of capture. Organizations can accomplish this solution by configuring the different capture mechanisms to filter out traffic based on given characteristics.

Organizations should not blindly decide what traffic to filter out. They must evaluate what type of traffic will give the most benefit for the available assets. In this report, we discussed different aspects of traffic that may be important when engaging in such an evaluation along with a suggested method for defining how critical different traffic is to investigations and how much risk storing that traffic brings to the organization.

The provided examples show how traffic volumes break down in several networks and grow over time. These volumes show how storage requirements change based on what traffic is stored in which traffic capture tier. The examples also show how volumes change and why it is important to plan storage needs for several years down the road.

Appendix A: Detailed Ranking Charts

Applicable Captured Data Uses

Step 1: In the "Y/N" column, indicate whether or not each purpose for using traffic capture is applicable for your organization. The last three columns show the different information each storage tier can provide for each purpose.

Y/N	Reason Number	Purpose	Capture Method*	Who	How Much	When	How Long	Using What	Transfer-ring What	How
	1	Investigate attacks	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓	✓	✓	
__			P	✓	✓	✓	✓	✓	✓	✓
	2	Police policies	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓	✓	✓	
__			P	✓	✓	✓	✓	✓	✓	✓
	3	Provide information to create policies	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓	✓	✓	
__			P	✓	✓	✓	✓	✓	✓	✓
	4	Understand normal network traffic	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓			
__			P	✓	✓	✓	✓			
	5	Understand abnormal network traffic	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓			
__			P	✓	✓	✓	✓			
	6	Plan for network upgrades	N	✓	✓	✓	✓			
			A	✓	✓	✓	✓			
__			P	✓	✓	✓	✓			

*** N = network flow, A = augmented flow, P = pcap**

Desired Storage Days

Step 1: Enter the number of days that it would be most useful for your organization to have data for each of the purposes from the *Applicable Captured Data Uses* chart.

Capture Method	Selected Network Data Usage Purpose	Days of Traffic Needed	Days to Store
Pcap	1	To investigate attacks: ___	See Attack/Risk and Requirements Charts
	2	To police policies based on traffic content: ___	Consider using filtering/IDS
	3	To see what content is passing to create network use policies: ___	Consider using filtering/IDS
Augmented Flow	1	To investigate attacks: ___	See Attack/Risk and Requirements Charts
	2	To police policies based on traffic content: ___	Consider using filtering/IDS
	3	To see what content is passing to create network use policies: ___	Consider using filtering/IDS
Network Flow	1	To investigate attacks: ___	See Attack/Risk and Requirements Charts
	2	To see volume, who, and when information to police policies based on traffic content after the fact: ___	____
	3	To see volume, who, and when information to create network use policies: ___	____
	4, 5	To trend traffic to understand normal and abnormal traffic: ___	____
	6	To trend traffic to plan for upgrades: ___	____

Attack Type Criticality

Step 1: For each attack type, determine the ramifications of not being able to determine the tier-specific information if that attack occurs. For example, if the organization may have legal ramifications if it cannot determine how and what exact data was exfiltrated, enter 3 in the Pcap column of the "Exfiltration" row.

		Criticality of determining:		
		Network Flow (N)	Augmented Flow (A)	Pcap (P)
		who is affected, how much, when, how often	using what, transferring what	how and what data
Attack Category	Scanning			
	Indexing			
	Malware Drop			
	C2/Backdoor Communications			
	Worm Propagation			
	System Control			
	Exfiltration			
	Data Corruption			
	DoS/DDoS Flooding			
	DoS/DDoS Crashing			

Criticality Levels
3: Legal ramifications/major increase in financial losses
2: Increase in financial losses or time to resolve issues caused by the incident
1: None or minor inconvenience

Effectiveness of Having Attack Data

Step 1: Define the "Column Value" ranges as a set number of days. Use the upper value of the range as the column value. See the bottom row of the chart for effectiveness values. For example, organizations can use 1, 7, 28, 90, and 180 days for immediate, short-term, mid-term, long-term, and archive time ranges, respectively.

Step 2: Determine how effective having each attack type data is for each time period. For example, if exfiltration is not identified on average until three months after it occurred, the 90- and 180-day columns would be given a 5, and the other columns would be given lower values.

Step 3: Enter the column value for the column with the highest effectiveness value. Row values should trend up, trend down, or remain steady—there should not be multiple peaks. If multiple effectiveness time frames tie for highest effectiveness value, enter the column value that corresponds to the longest time frame.

	Column Value:	____	____	____	____	____	
	Effectiveness Time Range (in days):	**Immediate**	**Short-term**	**Mid-term**	**Long-term**	**Archive**	**Days of Greatest Effectiveness**
Attack Category	**Scanning**						
	Indexing						
	Malware Drop						
	C2/Backdoor communications						
	Worm Propagation						
	System Control						
	Exfiltration						
	Data Corruption						
	DoS/DDoS Flooding						
	DoS/DDoS Crashing						

Effectiveness Value

5: Five times the effectiveness of value 1—data at this time frame is critical
4: Four times the effectiveness of value 1
3: Three times the effectiveness of value 1
2: Two times the effectiveness of value 1
1: Data at this time frame has little use

Services to Filter

Step 1: In the *Risk and Requirements Chart*, black out the cells for combinations that are unlikely to be useful. In each remaining cell in the "Attack Category/Storage Tier" columns, enter the corresponding criticality value from the *Attack Type Criticality Chart*.

Step 2: Enter a risk value for each attack type at each storage tier in the "Risk Value" column.

Tier	Description
Network Flow	• Network structure
Augmented Flow	• PII • E-discovery (if storing embedded files only) • Malware (if storing embedded files only) • Business confidential data • Network structure
Pcap	• PII • E-discovery • Malware • Phone conversations (TCP: VoIP only) • Web cam content • Credentials (from unencrypted traffic) • Business confidential data • Network structure

Step 3: Calculate Essentiality as Max (Attack Type Criticality) – Risk Value for each attack at each storage tier.

Step 4: Enter Useful Days from Days of Greatest Effectiveness in "Effectiveness of Having Attack Data."

Step 5: Calculate storage per day for each level where the service will be stored. For metadata, add extra overall storage to the totals for extracted files if applicable. For pcap, add extra overall storage to the totals for the file format overhead.

$$N = \text{Flows} \times \text{Flow Record Size}$$

$$A = \text{Flows} \times \text{Augmented Flow Record Size}$$

$$P = ((\text{Bytes} / \text{File Size}) \times \text{Overhead}) + \text{Bytes}$$

Step 6: Sum the storage size values for each tier. If the values are too high, use Essentiality to determine which services to filter, starting with the services having the smallest values.

Traffic Type/Service	Scanning			Indexing			Malware Drop			C2/Backdoor Communications			Worm Propagation			System Control			Exfiltration			Data Corruption			DoS/DDoS Flooding			DoS/DDoS Crashing			Risk Value			Essentiality			Effectiveness	Max Expected Traffic per Day		Storage per Day		
	N	A	P	N	A	P	N	A	P	N	A	P	N	A	P	N	A	P	N	A	P	N	A	P	N	A	P	N	A	P	N	A	P	N	A	P	Useful Days	Bytes	Flows	N	A	P
TCP: HTTP																																										
TCP: Encrypted HTTP																																										
TCP: Remote Connections																																										
TCP: Encrypted Tunneling																																										
TCP: Email																																										
TCP: File Copy																																										
TCP: Encrypted File Copy																																										
TCP: VoIP																																										
TCP: Encrypted VoIP																																										
TCP: Generic (all other TCP)																																										
UDP: DNS																																										
UDP: NTP																																										
UDP: Remote Connections																																										
UDP: Encrypted Tunneling																																										
UDP: VoIP																																										
UDP: Encrypted VoIP																																										
UDP: Generic (all other UDP)																																										
ICMP																																										
ESP																																										
IPv4																																										
IPv6																																										
Other																																										
																																					Totals:					
																																					Storage Days Goal:					

N = network flow, A = augmented flow, P = pcap

Risk Values

5: Storage of information has legal ramifications that cannot be mitigated.

4: Storage or leakage of information has legal or major financial ramifications that can be mitigated, but mitigation is not already in place.

3: Storage or leakage of information has financial ramifications that can be mitigated, but mitigation is not already in place.

2: Storage or leakage of information has legal or financial ramifications, but mitigations are in place.

1: Storage or leakage of information would cause minor inconvenience. This value is used anywhere the ramifications are unlikely to be anything other than leaked network structure if the stored data was subject to unauthorized access.

Appendix B: Process for Augmented Packet Capture

Appendix B provides a flow chart that outlines the methodology presented in Section 4. This flow chart will help organizations understand the steps required to rank and capture network data flows at the correct tiers and for the right amount of time.

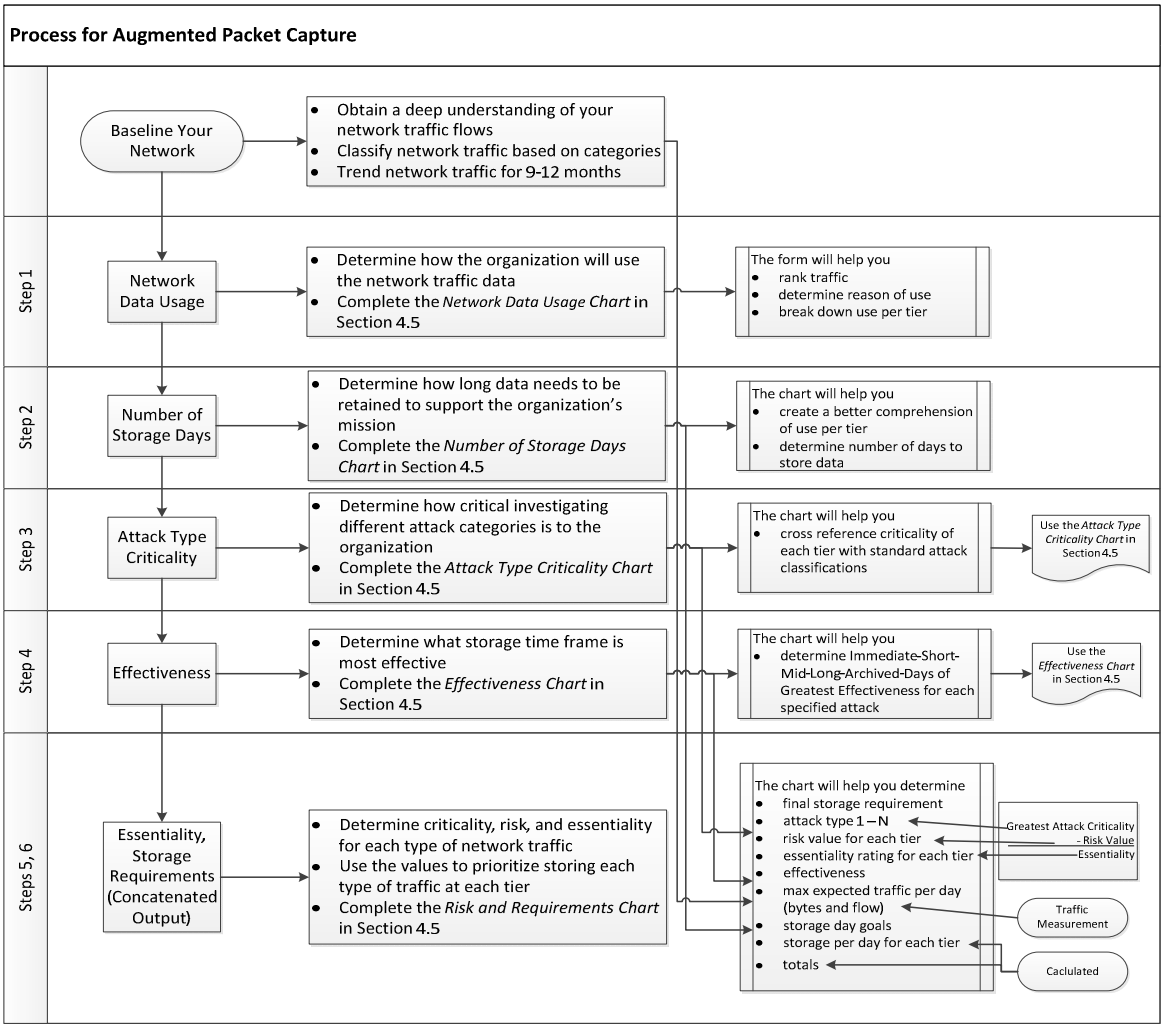

Figure 5: Smart Collection and Storage Process

References

[Aceto 2013a]
Aceto, G.; Botta, A.; Pescape, A.; & Westphal, C. “Efficient Storage and Processing of High-Volume Network Monitoring Data.” *IEEE Transactions on Network and Service Management 10*, 2 (June 2013): 162-175.

[Aceto 2013b]
Aceto, G.; Botta, A.; de Donato, W.; & Pescape, A. “Cloud Monitoring: A Survey.” *Computer Networks 57*, 9 (June 19, 2013): 2,093-2,115.

[Banks 2013]
Banks, D. “Custom Full Packet Capture System,” SANS, 2013.

[Bejtlich 2013]
R. Bejtlich. *The Practice of Network Security Monitoring Understanding Incident Detection and Response.* No Starch Press, 2013.

[Chou 2013]
Chou, T. “The Truth About Technology Costs.” *CFO* (February 6, 2013). http://ww2.cfo.com/the-cloud/2013/02/the-truth-about-technology-costs/

[Claise 2013]
Claise, B. (ed.); Trammell, B. (ed.); & Aitken, P. *Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information*. Internet Engineering Task Force (IETF), 2013. http://tools.ietf.org/html/rfc7011

[DE-CIX 2014]
DE-CIX. *Statistics*. https://www.de-cix.net/about/statistics/ (2014).

[Delcroix 2013]
Delcroix, J.-C.; Keene, I.; Petri, G.; Rickard N.; & Tian, T. “How Cloud, Mobile and Video Will Increase Enterprise Bandwidth Needs Through 2017.” *Gartner* (March 22, 2013). https://www.gartner.com/doc/2383515/cloud-mobile-video-increase-enterprise

[Deri 2013]
Deri, L.; Cardigliano, A.; & Fusco, F. “10 Gbit Line Rate Packet-to-Disk Using n2disk” 3,399-3,404. *2013 Proceedings IEEE INFOCOM*. Turin, Italy, Apr. IEEE, 2013.

[European Information Security Summit 2014]
European Information Security Summit. *Emulex Announces Next Generation of High Performance PCI Express 3.0 EndaceDAG Data Capture Cards for Network Monitoring Applications*. RealWire (February 18, 2014). http://www.realwire.com/releases/Emulex-High-Performance-PCI-Express-30-EndaceDAG-Data-Capture-Cards

[Francois 2013]
Francois, J.; State, R.; & Engel, T. “Aggregated Representations and Metrics for Scalable Flow Analysis” 478-482. *2013 IEEE Conference on Communications and Network Security (CNS).* Washington, D.C., Oct. 2013. IEEE, 2013.

[Gupta 2012]
Gupta, S. *Logging and Monitoring to Detect Network Intrusions and Compliance Violations in the Environment*. SANS, 2012.

[Hutchins 2011]
Hutchins, E. M.; Cloppert, M. J.; & Amin, R. M. *Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains*. Lockheed Martin, 2011.

[Information Security Forum 2007]
Information Security Forum. *The Standard of Good Practice for Information Security*. ISF, 2007.

[ISACA 2012]
ISACA. *Cobit 5: A Business Framework for the Governance and Management of Enterprise IT*. ISACA, 2012.

[ISO 2009]
International Organization for Standardization (ISO). *ISO/IEC 27033-1:2009, Information Technology -- Security Techniques -- Network Security*. ISO, 2009.

[Mandiant 2013]
Mandiant. *APT 1 Exposing One of China's Cyber Espionage Units*. Mandiant, 2013.

[Mannie 2004]
Mannie, E., ed. *Generalized Multi-Protocol Label Switching (GMPLS) Architecture*. Internet Engineering Task Force (IETF), 2004. http://tools.ietf.org/html/rfc3945

[McCanne 1992]
McCanne, S. & Jacobson, V. *The BSB Packet Filter: A New Architecture for User-level Packet Capture*. Lawrence Berkeley Laboratory, 1992. http://www.tcpdump.org/papers/bpf-usenix93.pdf

[Moyle 2012]
Moyle, E. & Kelley, D. “Network Monitoring as a Security Tool.” *InformationWeek*, Sep. 30, 2012. http://reports.informationweek.com/abstract/21/9048/Security/
Strategy:-Network-Monitoring-as-a-Security-Tool.html

[Napatech 2014]
Napatech. *Napatech NT Network Adapter with Napatech Software Suite Filtering User's Manual*. Napatech, 2014.

[Netflix 2014]
Netflix. *Manage Bandwidth Usage*. https://support.netflix.com/en/node/87 (2014).

[NISO 2004]
National Information Standards Organization (NISO). *Understanding Metadata*. NISO, 2004.

[NIST 2011]
National Institute of Standards and Technology (NIST). *NIST Special Publication 800-137: Information Security Continuous Monitoring (ISCM) for Federal Information Systems and Organizations*. NIST, 2011. http://www.nist.gov/manuscript-publication-search.cfm?pub_id=909726

[NIST n.d.]
National Institute of Standards and Technology (NIST). *International System of Units (SI): Prefixes for Binary Multiples*. http://physics.nist.gov/cuu/Units/binary.html (no date).

[PCI 2013]
PCI Security Standards Council. *Payment Card Industry Data Security Standard (PCI DSS)*. PCI Security Standards Council, 2013. https://www.pcisecuritystandards.org/security_standards/documents.php?document=pci_dss_v2-0#pci_dss_v2-0

[Powers 2010]
Powers, A. "Tuning Cisco's Flexible NetFlow 'Flow Record' Definitions to Conserve Bandwidth." *NetFlow Ninjas Blog* (January 20, 2010). http://www.lancope.com/blog/tuning-your-netflow-v9-record-definitions-to-conserve-bandwidth/

[Quittek 2004]
Quittek, J.; Zseby, T.; Claise, B.; & Zander, S. *RFC 3917: Requirements for IP Flow Information Export (IPFIX)*. Internet Engineering Task Force (IETF), 2004. http://tools.ietf.org/html/rfc3917

[Riverbed Technology 2013]
Riverbed Technology. *WinPcap*. http://www.winpcap.org/ (2013).

[Sanders 2013]
Sanders, C. & Smith, J. *Applied Network Security Monitoring: Collection, Detection, and Analysis*. Syngress, 2013.

[SEI 2014]
Software Engineering Institute, Carnegie Mellon University. *SiLK FAQ*. https://tools.netsa.cert.org/silk/faq.html (2014).

[Shimeall 2010]
Shimeall, T.; Faber, S.; DeShon, M.; & Kompanek, A. *Using SiLK for Network Traffic Analysis Analyst's Handbook*. Software Engineering Institute, Carnegie Mellon University, 2010. http://tools.netsa.cert.org/silk/analysis-handbook.pdf

[Verizon 2013]
Verizon RISK Team. *2013 Data Breach Investigations Report*. Verizon, 2013.

[Winston 2004]
Winston, W. L. *Modeling Exponential Growth*. Microsoft Corporation, 2004.

[Wireshark 2013]
Wireshark. *Libpcap File Format*. http://wiki.wireshark.org/Development/LibpcapFileFormat (2013).

[Zseby 2009]
Zseby, T.; Molina, M.; Duffield, N.; Niccolini S.; & Raspall, F. *RFC 5475: Sampling and Filtering Techniques for IP Packet Selection*. IETF, 2009.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
(Leave Blank)	September 2014	Final

4. TITLE AND SUBTITLE	5. FUNDING NUMBERS
Smart Collection and Storage Method for Network Traffic Data	FA8721-05-C-0003

6. AUTHOR(S)

Angela Horneman and Nathan Dell

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Software Engineering Institute Carnegie Mellon University Pittsburgh, PA 15213	CMU/SEI-2014-TR-011

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFLCMC/PZE/Hanscom Enterprise Acquisition Division 20 Schilling Circle Building 1305 Hanscom AFB, MA 01731-2116	

11. SUPPLEMENTARY NOTES

12A DISTRIBUTION/AVAILABILITY STATEMENT	12B DISTRIBUTION CODE
Unclassified/Unlimited, DTIC, NTIS	

13. ABSTRACT (MAXIMUM 200 WORDS)

Captured network data enables an organization to perform routine tasks such as network situational awareness and incident response to security alerts. The process of capturing, storing, and evaluating network traffic as part of monitoring is an increasingly complex and critical problem. With high-speed networks and ever-increasing network traffic volumes, full-packet traffic capture solutions can require petabytes of storage for a single day. The capacity needed to store full-packet captures for a time frame that permits the needed analysis is unattainable for many organizations. A tiered network storage solution, which stores only the most critical or effective types of traffic in full-packet captures and the rest as summary data, can help organizations mitigate the storage issues while providing the detailed information they need. This report discusses considerations and decisions to be made when designing a tiered network data storage solution. It includes a method, based on a cost-effectiveness model, that can help organizations decide what types of network traffic to store at each storage tier. The report also uses real-world network measurements to show how storage requirements change based on what traffic is stored in which storage tier.

14. SUBJECT TERMS	15. NUMBER OF PAGES
network traffic, storage, filtering, improving traffic capture storage, traffic capture, planning, efficiency	75

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL