Privacy Preserving Network Flow Recording

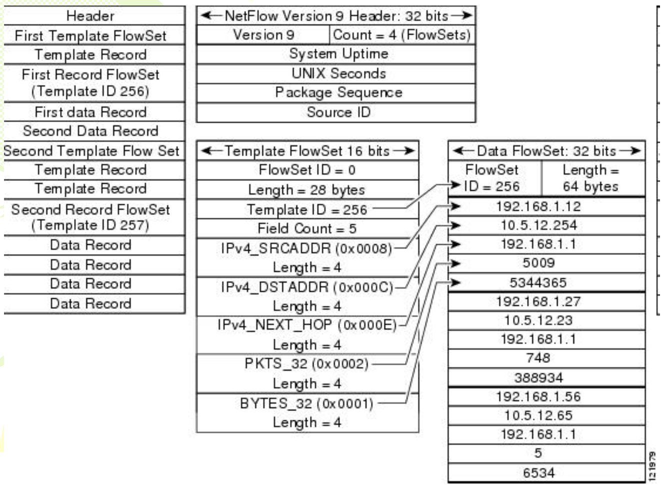
Bilal Shebaro (Computer Science-UNM)

Jedidiah R. Crandall (Computer Science-UNM)

The University of New Mexico

Basic Idea

- Most ISPs and institutions use NetFlow
- NetFlow records are stored in plain most of the time
- Websites, webservices & applications have signatures
- We implemented a privacy preserving way of storing NetFlow records and generating statistical reports
  - IBE & P.P. semantics for on-the-fly statistics



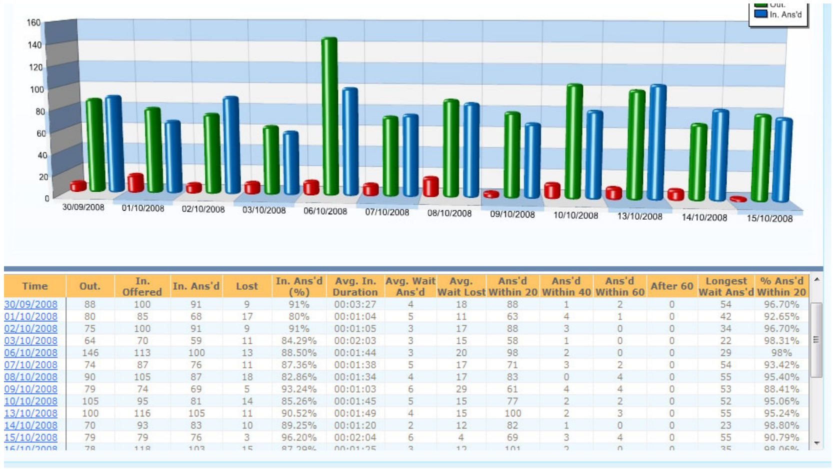
Header
First Template FlowSet
Template Record
First Record FlowSet (Template ID 256)
First data Record
Second Data Record
Second Template Flow Set
Template Record
Template Record
Second Record FlowSet (Template ID 257)
Data Record
Data Record
Data Record
Data Record

NetFlow Version 9 Header: 32 bits
Version 9 | Count = 4 (FlowSets)
System Uptime
UNIX Seconds
Package Sequence
Source ID

Template FlowSet 16 bits
FlowSet ID = 0
Length = 28 bytes
Template ID = 256
Field Count = 5
IPv4_SRCADDR (0x0008)
Length = 4
IPv4_DSTADDR (0x000C)
Length = 4
IPv4_NEXT_HOP (0x000E)
Length = 4
PKTS_32 (0x0002)
Length = 4
BYTES_32 (0x0001)
Length = 4

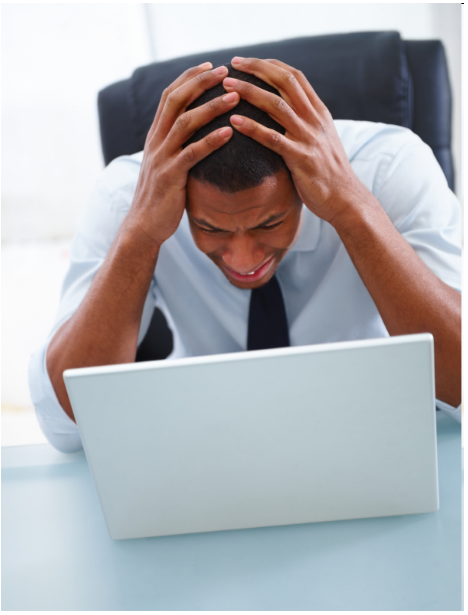
Data FlowSet: 32 bits
FlowSet ID = 256 | Length = 64 bytes
192.168.1.12
10.5.12.254
192.168.1.1
5009
5344365
192.168.1.27
10.5.12.23
192.168.1.1
748
388934
192.168.1.56
10.5.12.65
192.168.1.1
5
6534

NetFlow Records

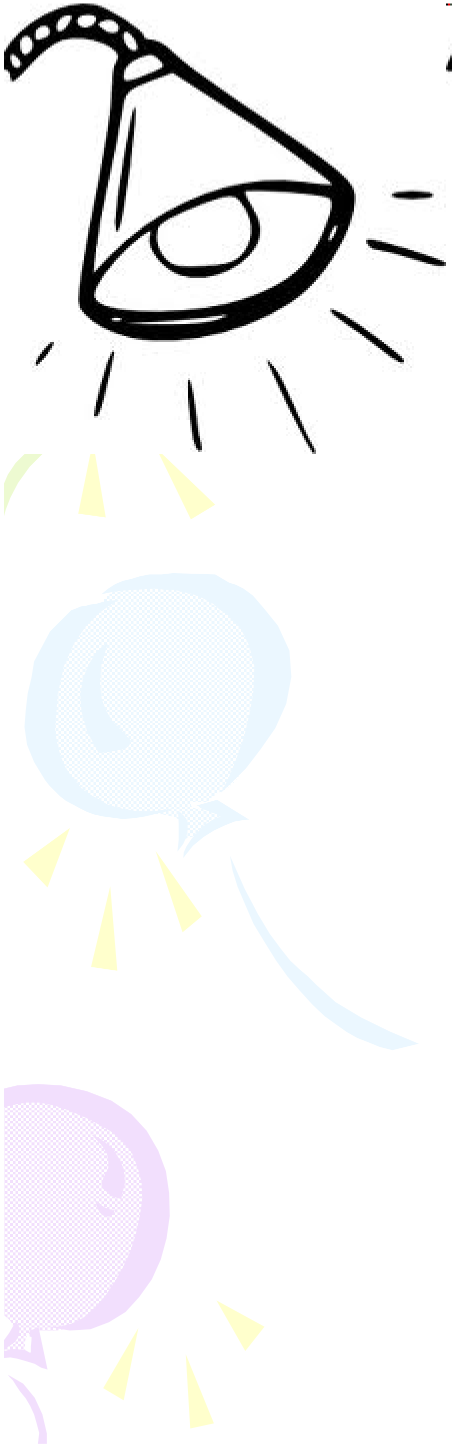


Time	Out.	In. Offered	In. Ans'd	Lost	In. Ans'd (%)	Avg. In. Duration	Avg. Wait Ans'd	Avg. Wait Lost	Ans'd Within 20	Ans'd Within 40	Ans'd Within 60	After 60	Longest Wait Ans'd	% Ans'd Within 20
30/09/2008	88	100	91	9	91%	00:03:27	4	18	88	1	2	0	54	96.70%
01/10/2008	80	85	68	17	80%	00:01:04	5	11	63	4	1	0	42	92.65%
02/10/2008	75	100	91	9	91%	00:01:05	3	17	88	3	0	0	34	96.70%
03/10/2008	64	70	59	11	84.29%	00:02:03	3	15	58	1	0	0	22	98.31%
06/10/2008	146	113	100	13	88.50%	00:01:44	3	20	98	2	0	0	29	98%
07/10/2008	74	87	76	11	87.36%	00:01:38	5	17	71	3	2	0	54	93.42%
08/10/2008	90	105	87	18	82.86%	00:01:34	4	17	83	0	4	0	55	95.40%
09/10/2008	79	74	69	5	93.24%	00:01:03	6	29	61	4	4	0	53	88.41%
10/10/2008	105	95	81	14	85.26%	00:01:45	5	15	77	2	2	0	52	95.06%
13/10/2008	100	116	105	11	90.52%	00:01:49	4	15	100	2	3	0	55	95.24%
14/10/2008	70	93	83	10	89.25%	00:01:20	2	12	82	1	0	0	23	98.80%
15/10/2008	79	79	76	3	96.20%	00:02:04	6	4	69	3	4	0	55	90.79%
16/10/2008	78	118	103	15	87.29%	00:01:25	3	12	101	2	0	0	35	98.06%

Statistical Reports



Websites, Services, Web Applications, etc…

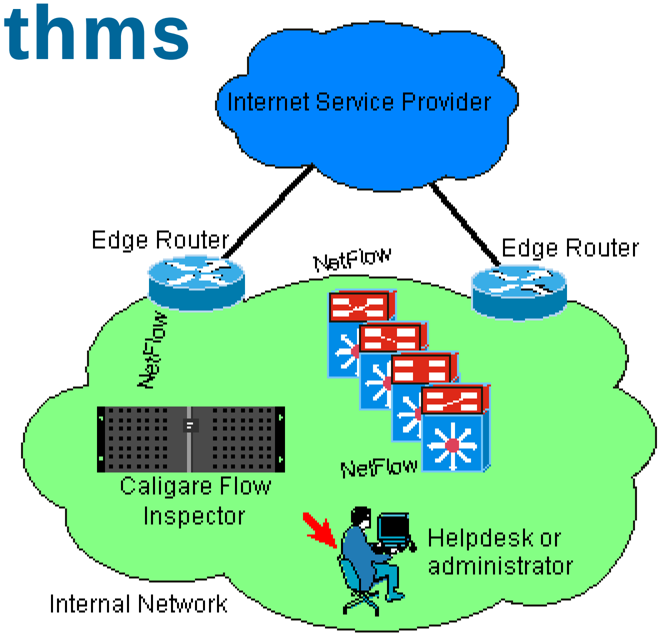


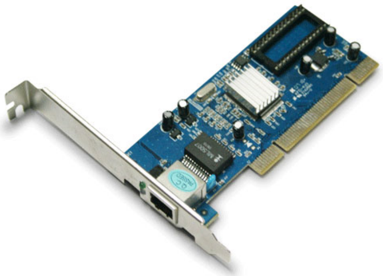
Outline

- Basic Idea
- Requirements
- NetFlow
- Threat Model and Challenges
- Scenarios
- Algorithm Steps, Queries, Setup
- Results
- Discussion and Future Work

Requirements

- Uses of NetFlow
- User interfaces for /20, /22, /24
- Network Traffic Generators & TCP-replay
- 3 Gbps Network Interface (tuntap)
- IBE + AES Encryption Algorithms
- Privacy Preserving Queries





NetFlow

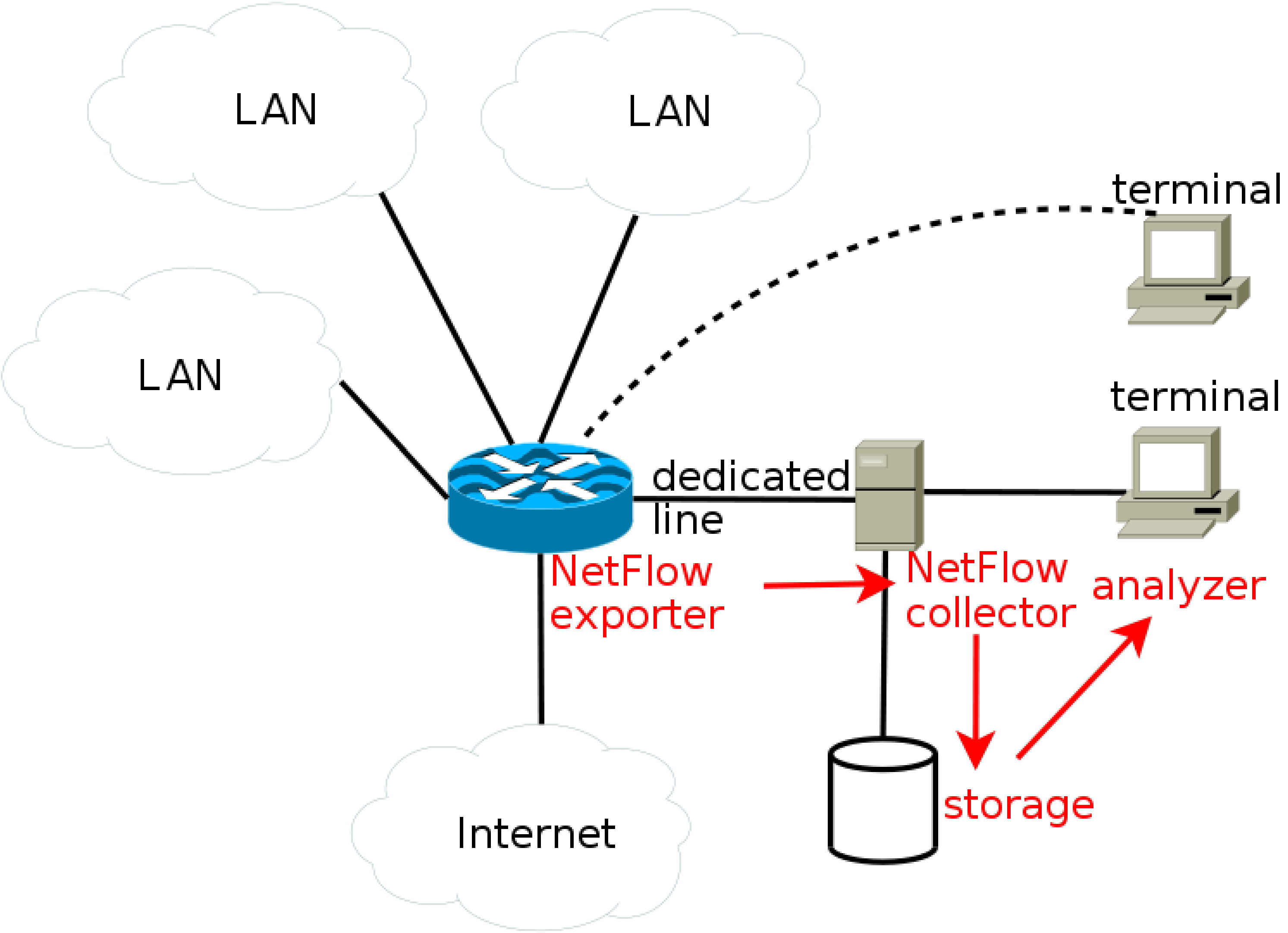
- Network protocol developed by Cisco Systems for collecting IP traffic information
  - Data recorded for the sake of network monitoring, traffic accounting, billing, network planning, security, DOS, etc...
  - Platforms supported: Cisco IOS, NXOS such as Juniper routers, Enterasys Switches, Linux, FreeBSD, NetBSD and OpenBSD.

- Version 5 and version 9 most popular

NetFlow

Sampled NetFlow

- rather than looking at every packet to maintain NetFlow records, the router looks at every *nth* packet
- Netflow version 5 have same sampling rate for all interfaces
- Netflow version 9 have different sampling rate per interface


LAN
LAN
LAN
terminal
terminal
dedicated line
NetFlow exporter
NetFlow collector
analyzer
storage
Internet

Traditional Cisco 7-tuple key Definition

1. Source IP address

2. Destination IP address

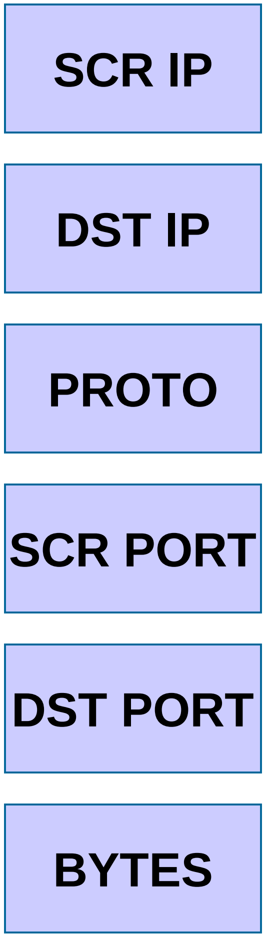
3. Source port for UDP or TCP

4. Destination port for UDP or TCP

5. IP protocol

6. Ingress interface (SNMP ifIndex)

7. IP Type of Service



Threat Model & Challenges

- NetFlow records in plain leaks confidential and individuals' private data
- Keep NetFlow recording useful in its all features
- Be able to generate useful statistical reports
- Leaving a security backdoor What's wrong with you???
- Recording, encryption and statistics data generated on the fly

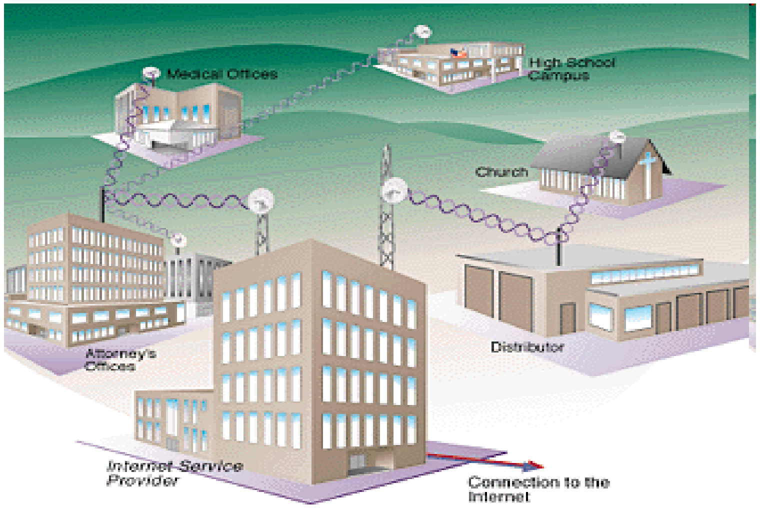
Threat Model & Challenges

- Forward & Backward Security
- Encrypt network flow data in privacy preserving way with no complicated public key infrastructure (IBE)
  - IP address + timestamp = public key
  - Decryption secret is not stored where encrypted data is stored
- Not all information could be encrypted
  - Statistical data
  - Privacy preserving semantics for DB

Scenario

- U.S. universities
- Network flow data is gathered for network management reasons
- State and federal law requires such data to be kept recorded for few weeks
- Breach of such information for employees is a privacy issue
- Our system supports both legal obligations and university network operations
- Decryption secret is distrubuted among:
  - Regents
  - Faculty senates
  - University council

Scenario



- ISPs

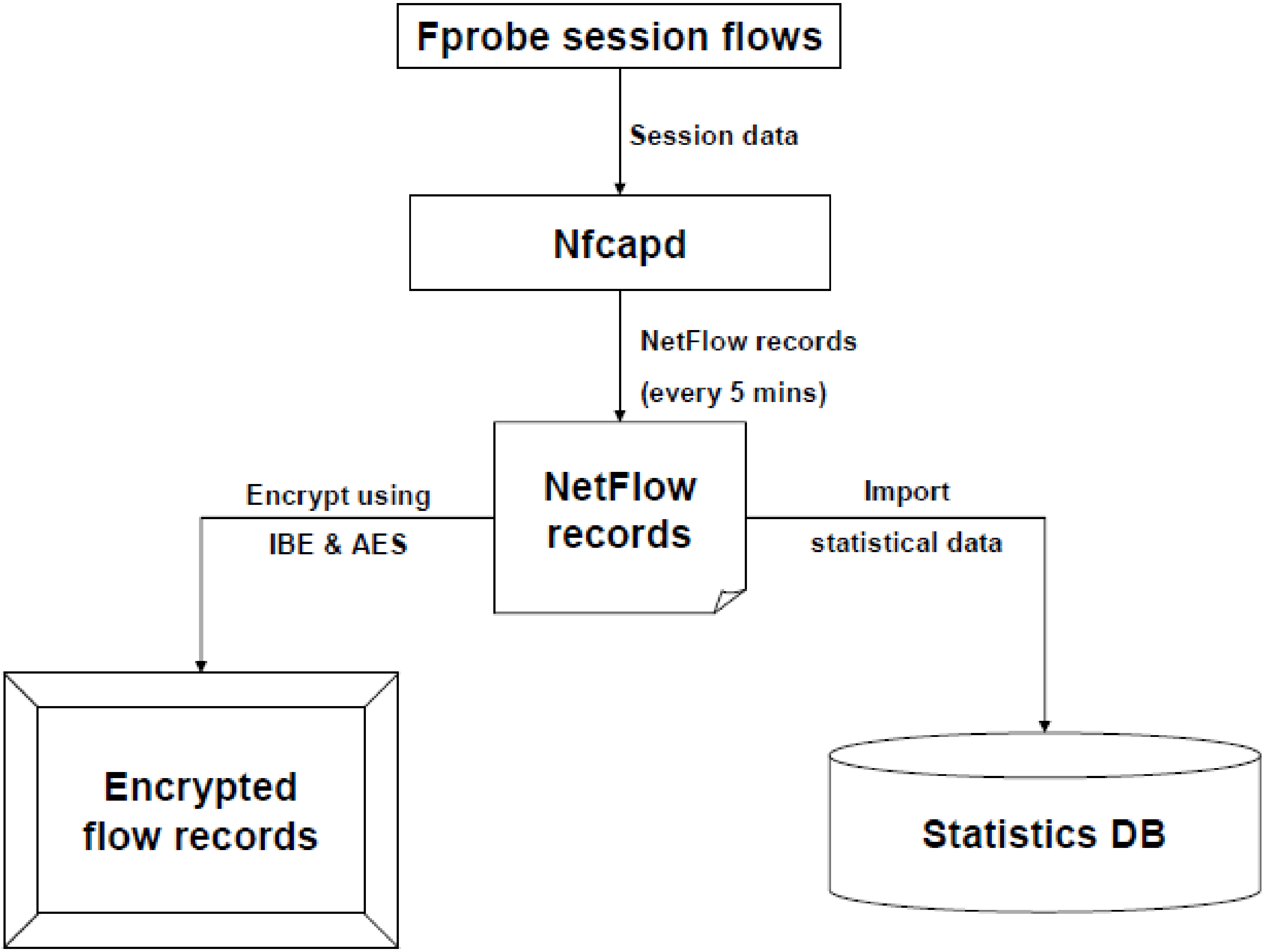
- Employees can access customers data to trace a network problem

- Decryption secret is distributed among:
  - Customer Service Department
  - Auditing department
  - Enforcing privacy policy organization

- We are NOT web privacy against untrusted network controllers

- We are making tools to enforce privacy policies so that network users could trust in network controllers

Big Picture



Fprobe session flows

Session data

Nfcapd

NetFlow records

(every 5 mins)

NetFlow records

Encrypt using

IBE & AES

Import

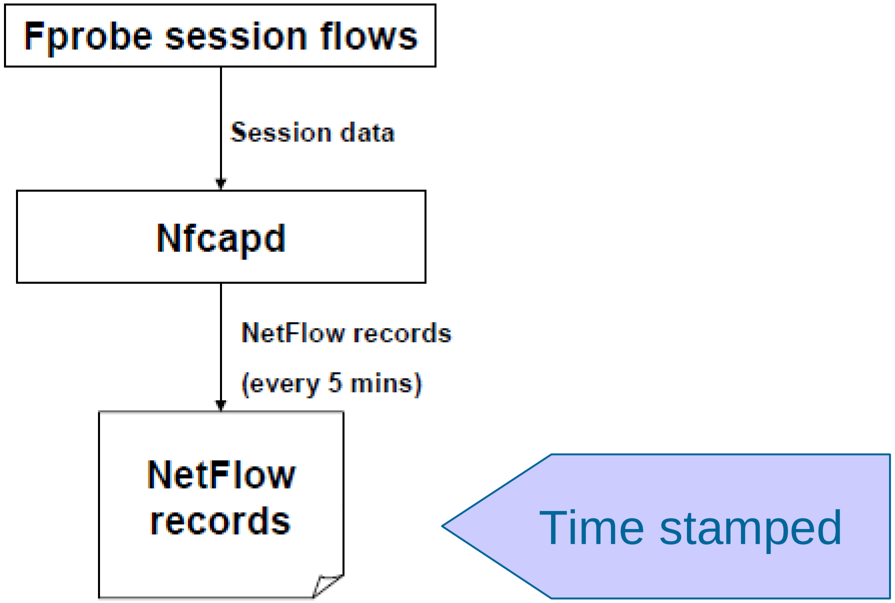
statistical data

Encrypted flow records

Statistics DB

Step 0: Data Collection

- Fprobe 1.1 running
- Nfcapd collects the flow and does file rotation every 5 minutes (configured)



Step 1: Flow Encryption

- Flows are combined per IP
- AES (128 key size) encrypts the flow
- IBE encrypts AES Key using:
  - Corresponding IP address
  - Corresponding file timestamp

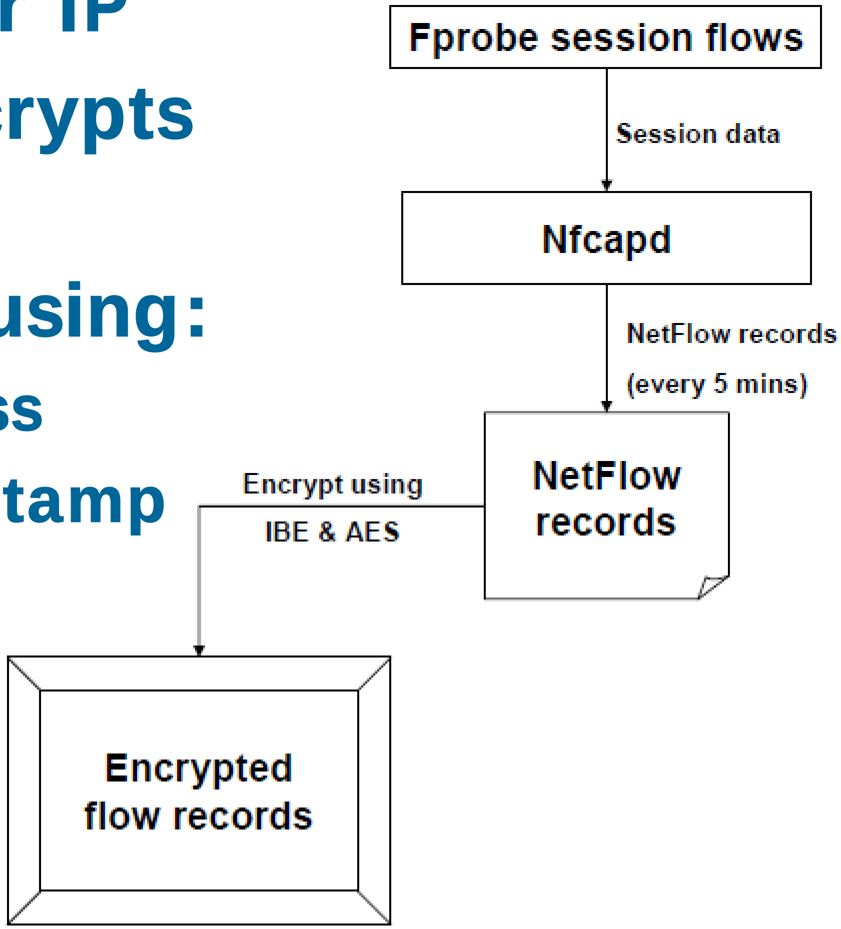
IP, IBE(AES-key), AES(flow record)

.

.

.

.



Fprobe session flows

Session data

Nfcapd

NetFlow records

(every 5 mins)

NetFlow records

Encrypt using

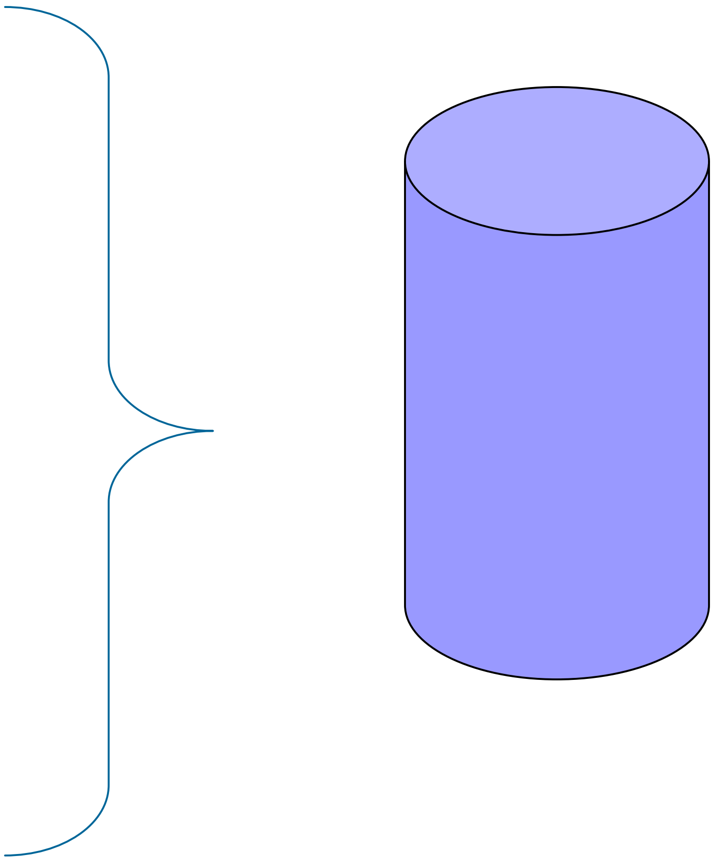
IBE & AES

Encrypted flow records

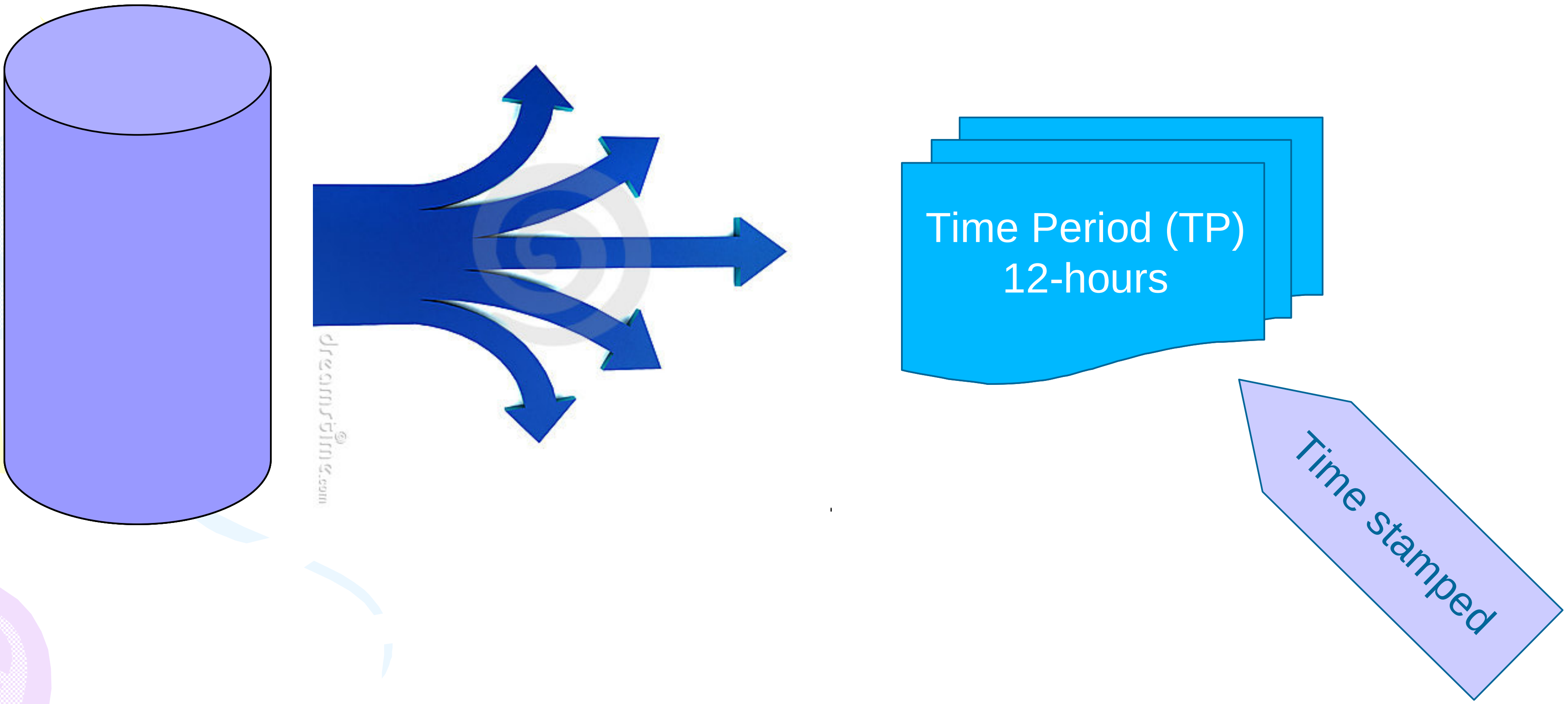
Step 2: Statistical Reports

- **Records are filtered out into:**

- IP Address
- TP: Time Period (time-stamped)
- TTI: Total TCP bytes In
- TTO: Total TCP bytes Out
- TUI: Total UDP bytes In
- TUO: Total UDP bytes Out
- LPI: List of Ports In
- LPO: List of Ports Out
- BI: Bytes In
- BO: Bytes Out
- PI: Packets In
- PO: Packets Out



Step 2: Statistical Reports



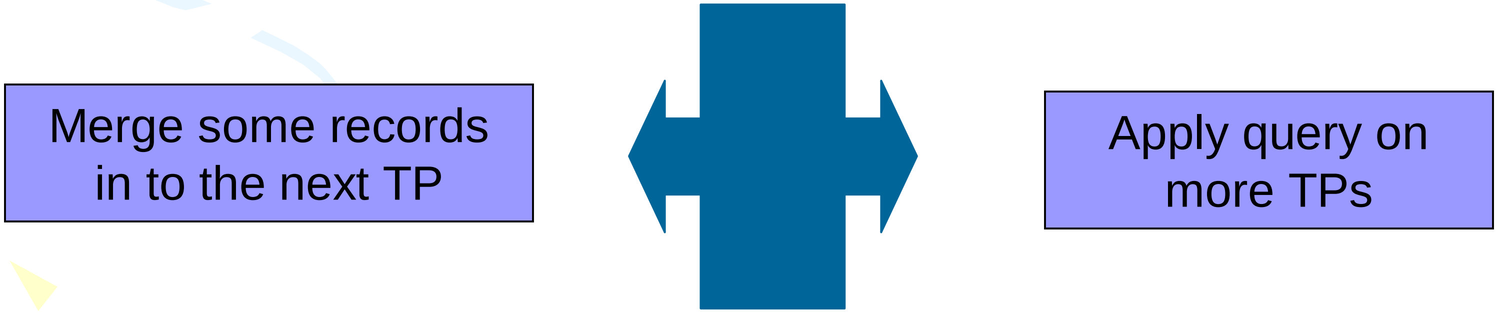
Time Period (TP)
12-hours

Time stamped

Step 2: Statistical Reports

- Reports require Queries
- Each Query has criteria and constraints
- Queries are applied on one or more TPs
- Queries applied on TPs that doesn’t match its criteria and constraints are rejected.

How to solve this:



Merge some records in to the next TP

Apply query on more TPs

Query Examples

(Link Utilization)

$$Q1 : Sum[BI, (TP \geq \alpha) \bullet IP]\ \ \&\ result \geq \beta$$

$$Q2 : Sum[BO, (TP \geq \alpha) \bullet IP]\ \ \&\ result \geq \beta$$

$$Q3 : Sum[BI + BO, (TP \geq \alpha) \bullet IP]\ \ \&\ result \geq \beta$$

Query Examples

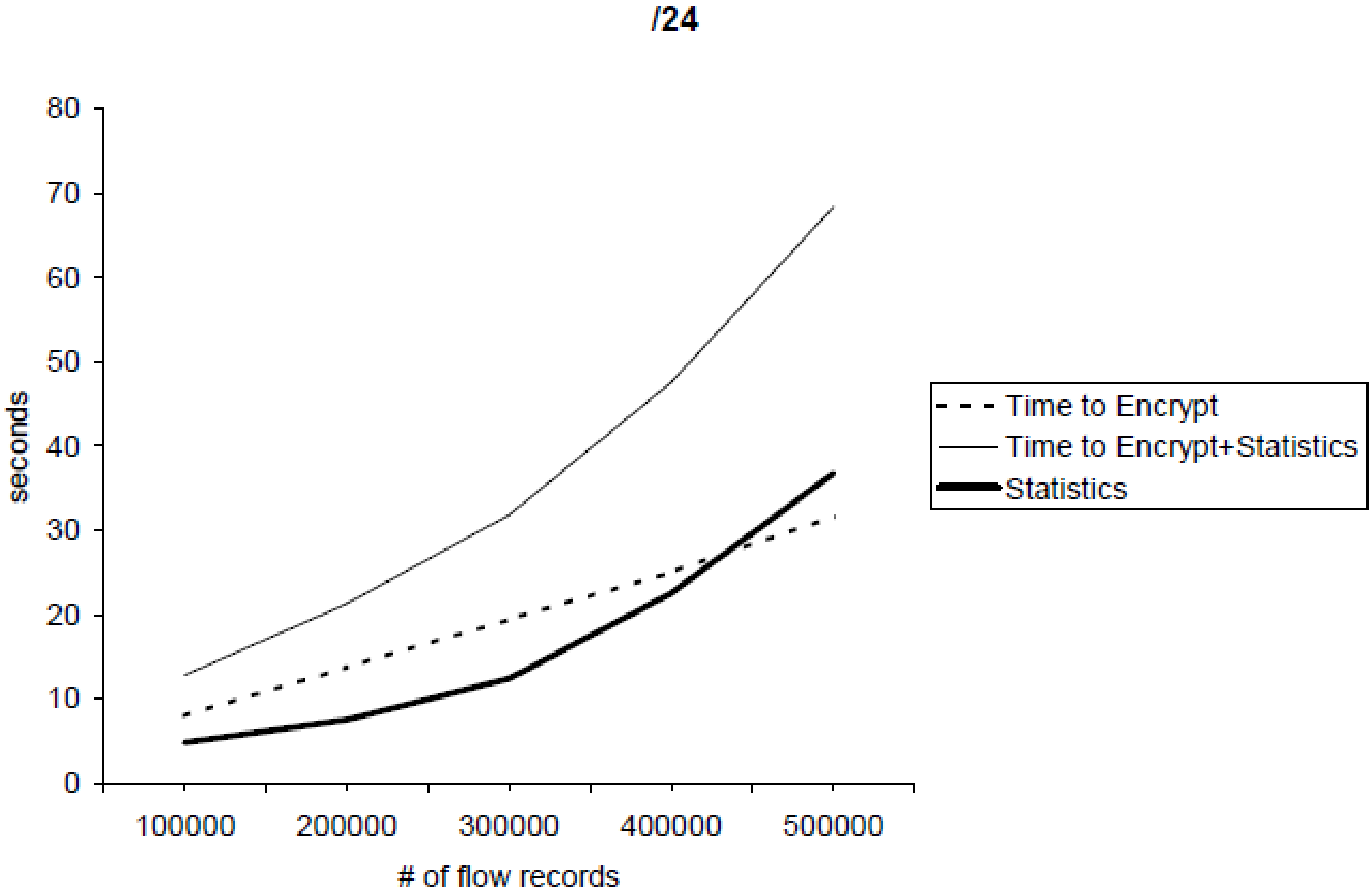
(Apps. Being used)

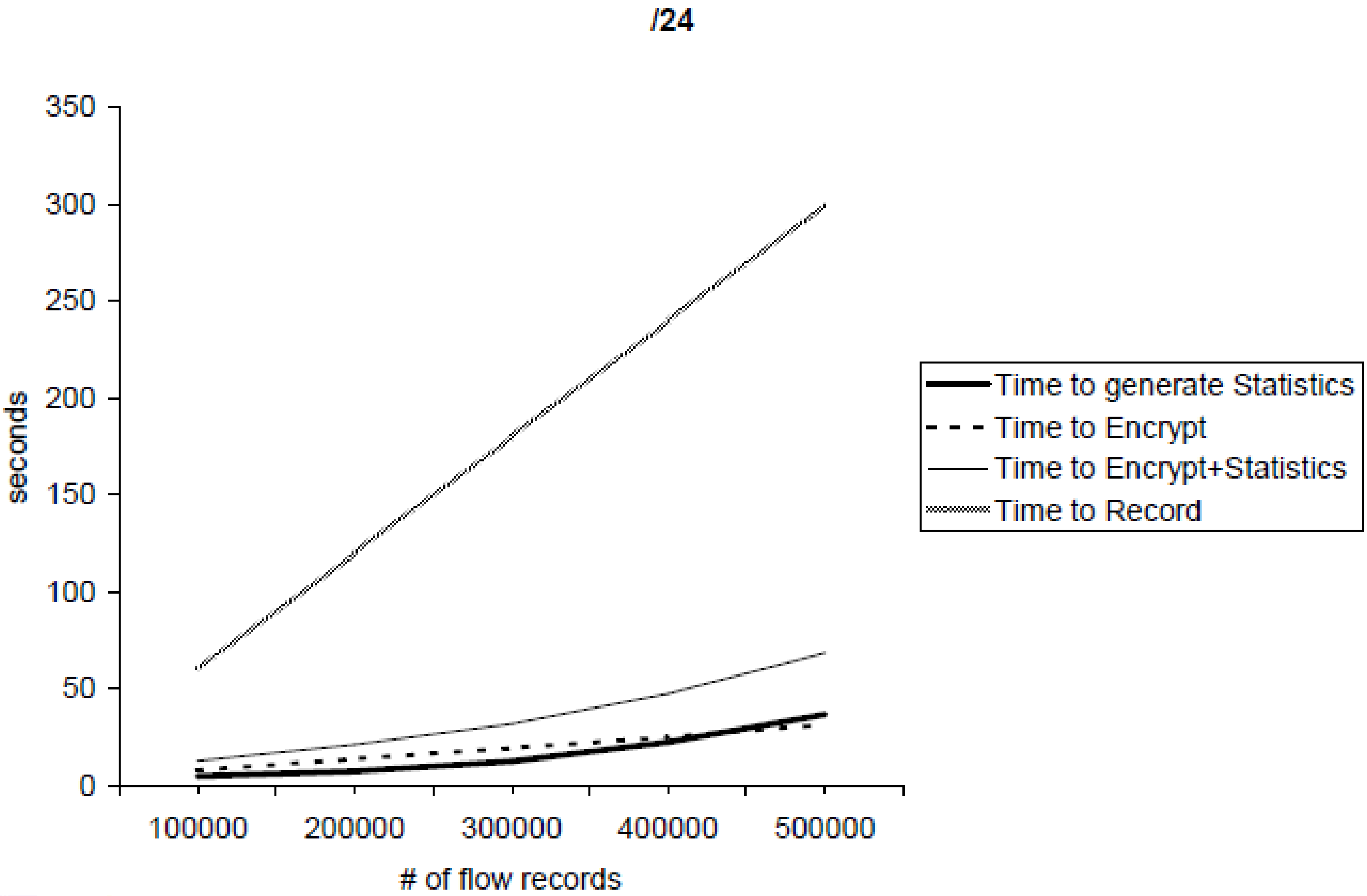
$$
\begin{aligned}
&Q5 : list[LPI, (TP \geq \alpha) \bullet IP_i] \\
+ \quad & list[LPO, (TP \geq \alpha) \bullet IP_i]
\end{aligned}
$$

$$\forall IP_i \in subnet,\ count(IP_{i_s}) > \delta$$

Setup

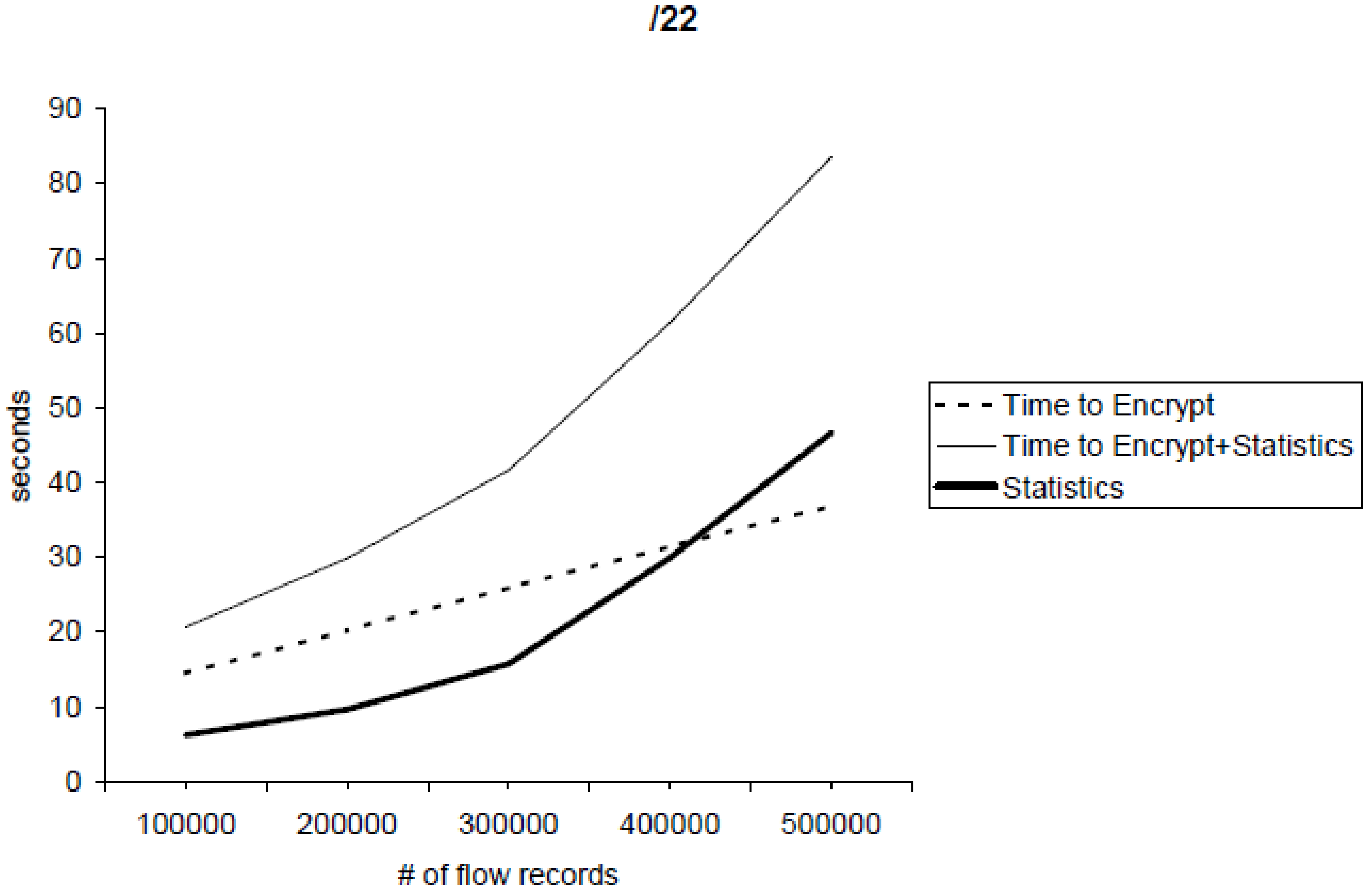
- /20, /22, /24 traffic data was generated.
- Core i7 X980 running at 3.33 GHz, 24 GB RAM, RAID 0 array with three 6 GB/s HD (motherboard RAID controller + PCI Express limited us to read at 3 Gbps from HD)
- Live capturing experiments for 6 hours for each subnet size (TCP-replay was used for that purpose)
- Measurements done for data recording, compared to encryption and statistical data importion

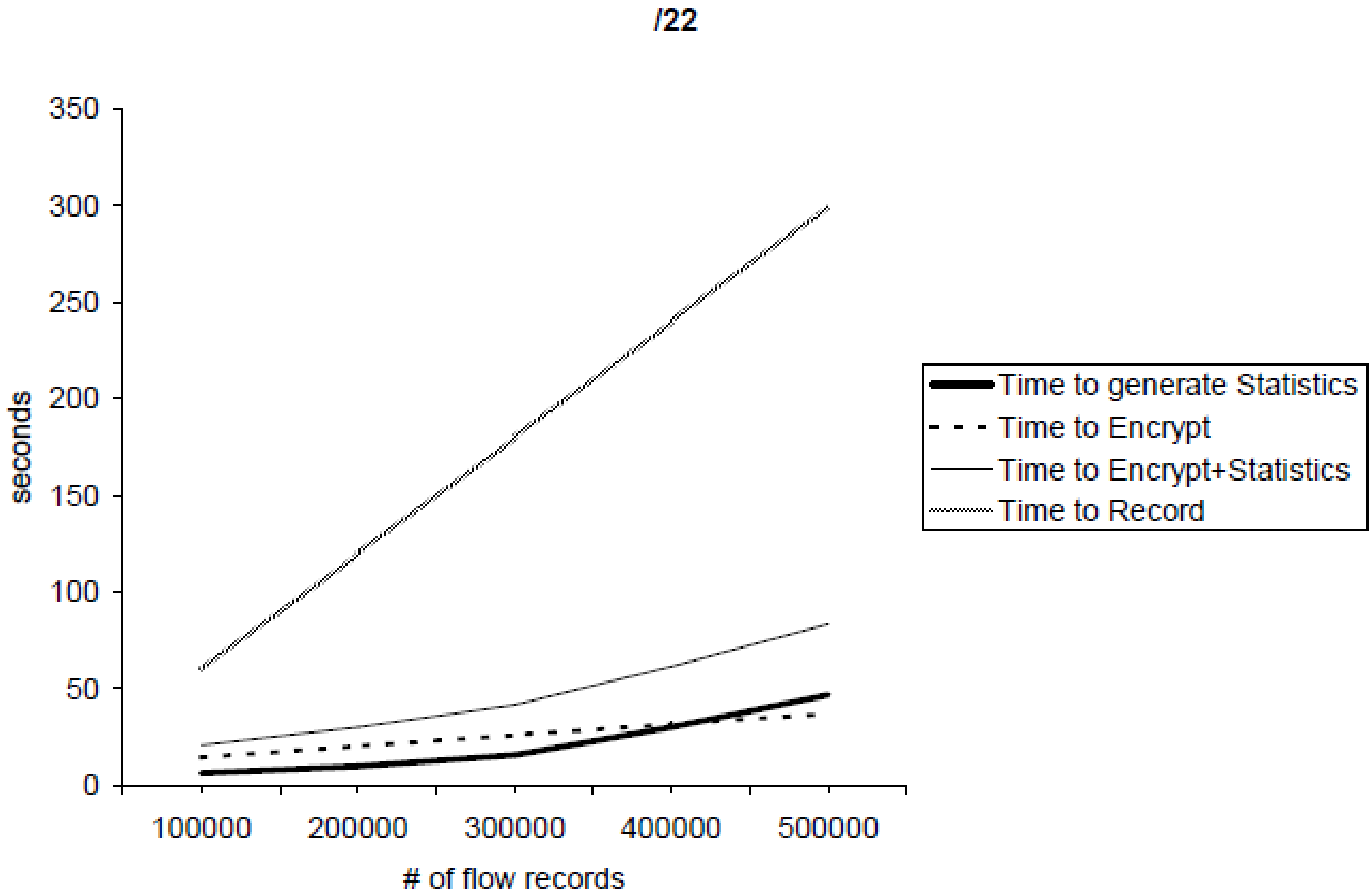

/24
80
70
60
50
40
30
20
10
0
seconds
100000
200000
300000
400000
500000
# of flow records
Time to Encrypt
Time to Encrypt+Statistics
Statistics

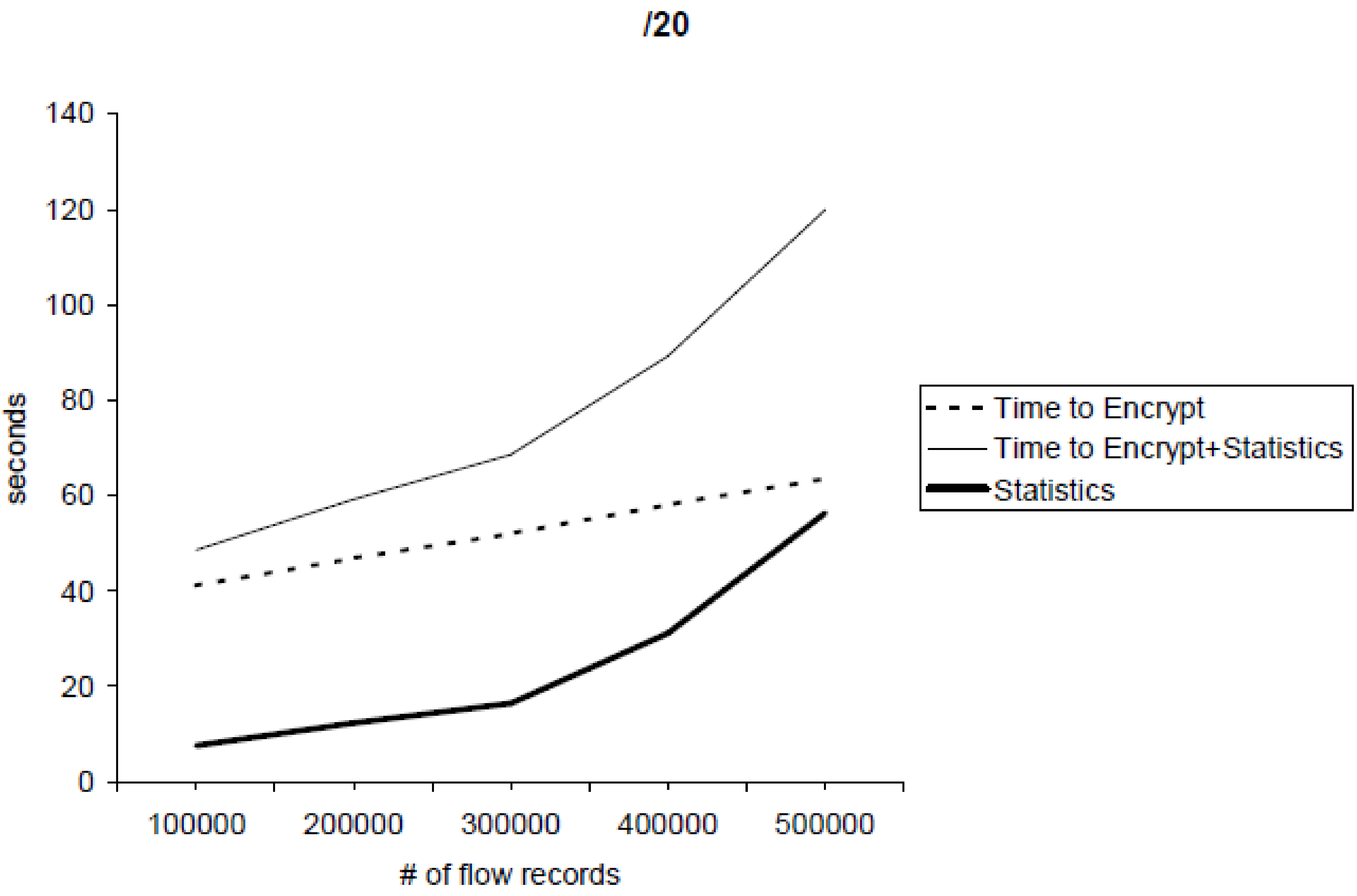


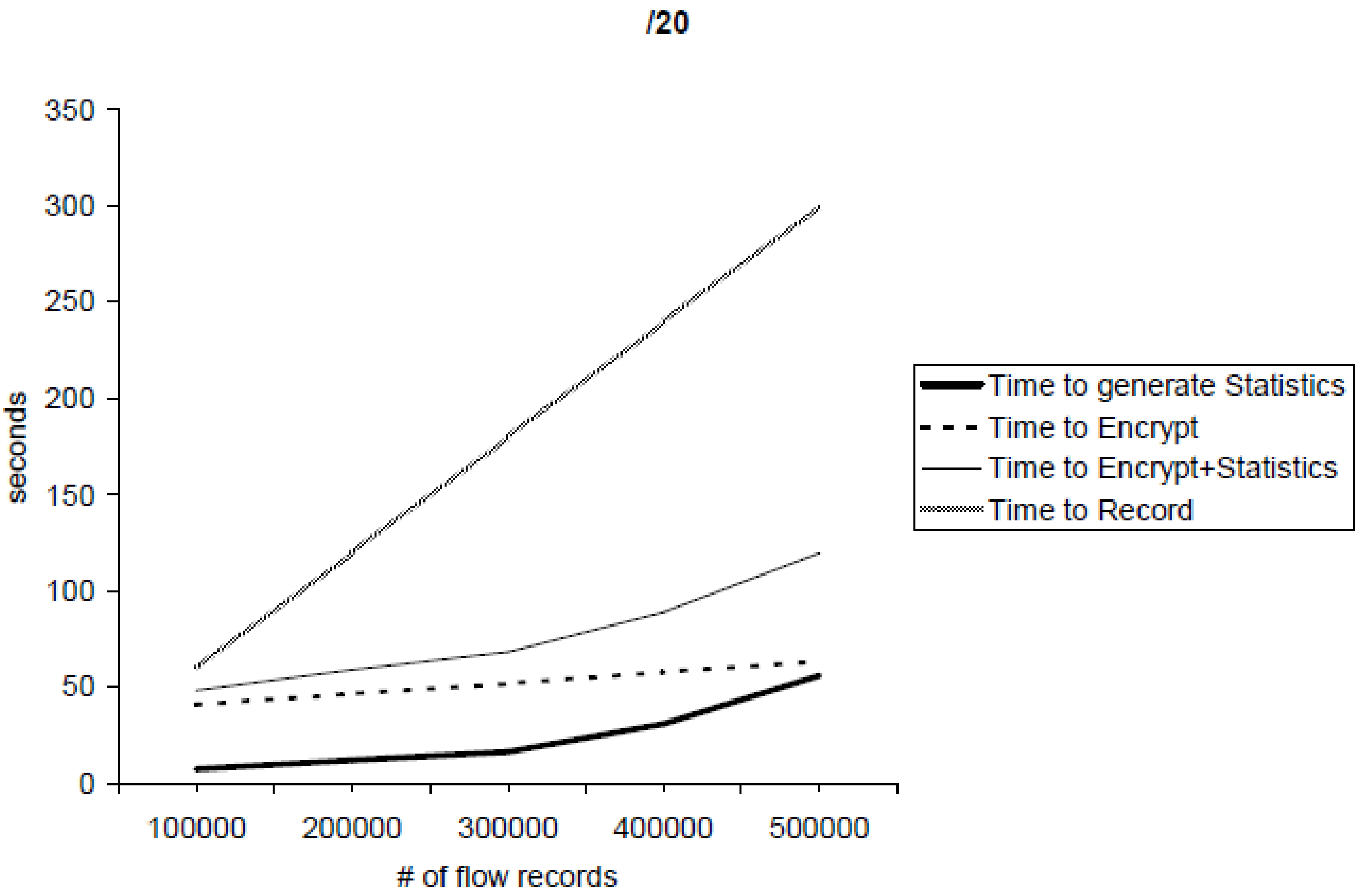
/24

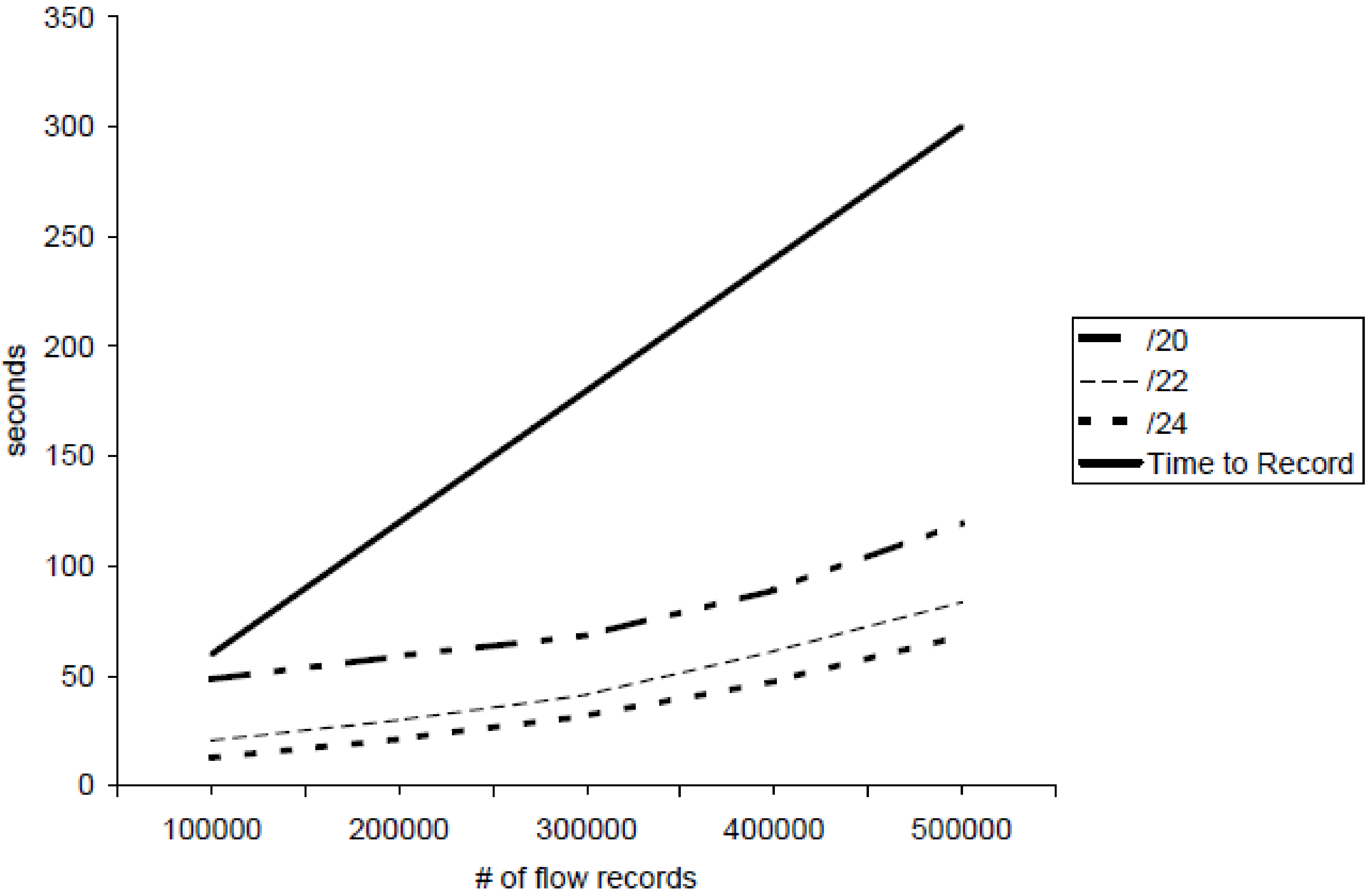
350
300
250
200
150
100
50
0
seconds
100000
200000
300000
400000
500000
# of flow records
Time to generate Statistics
Time to Encrypt
Time to Encrypt+Statistics
Time to Record


/22
seconds
90
80
70
60
50
40
30
20
10
0
100000
200000
300000
400000
500000
# of flow records
Time to Encrypt
Time to Encrypt+Statistics
Statistics


/22
350
300
250
200
150
100
50
0
seconds
100000
200000
300000
400000
500000
# of flow records
Time to generate Statistics
Time to Encrypt
Time to Encrypt+Statistics
Time to Record

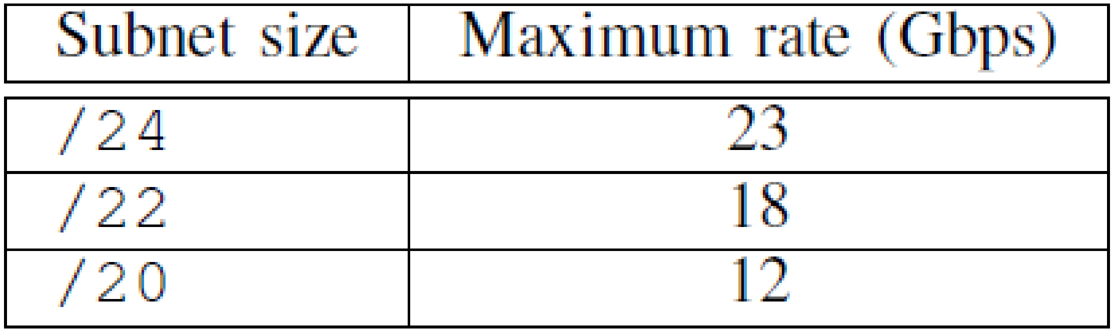

/20
140
120
100
80
60
40
20
0
seconds
100000
200000
300000
400000
500000
# of flow records
Time to Encrypt
Time to Encrypt+Statistics
Statistics


/20
seconds
350
300
250
200
150
100
50
0
100000
200000
300000
400000
500000
# of flow records
Time to generate Statistics
Time to Encrypt
Time to Encrypt+Statistics
Time to Record



350
300
250
200
150
100
50
0
seconds
100000
200000
300000
400000
500000
# of flow records
/20
/22
/24
Time to Record

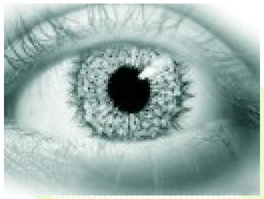
Offline Experiments

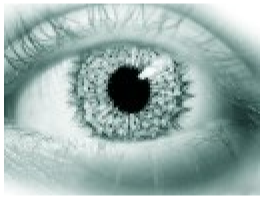


Subnet size	Maximum rate (Gbps)
/24	23
/22	18
/20	12

Discussion

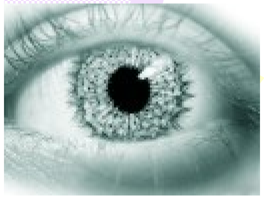
- Ability to encrypt + import statistical data within reasonable time
- Tradeoff in terms of how many distinct IP records need to be encrypted compared to indexing IP records in statistical DB
- Tradeoff between data accuracy and time intervals

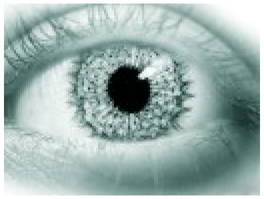




Future Work

- Better deal concerning the trade-offs
- Come up with a standard algorithm that can implement all kind of statistical queries
- Considering clickstream data to be stored in privacy preserving manner
- Tackle all network flow applications that records traffic and try to implement a privacy preserving version of them.





Acknowledgments

- NSF #0905177 & #0844880

"This material is based upon work supported by the National Science Foundation under Grant Nos. 0905177 and 0844880. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation."

NSF National Science Foundation
WHERE DISCOVERIES BEGIN

unm Computer SCIENCE