Traffic Clusters in Networks of Convenience

Ron McLeod, PhD. (Candidate)
Director - Research and Corporate Development
Telecom Applications Research Alliance
(TARA)

FloCon 2009

Who is TARA

- Private consortium of 35 member companies and research institutions all working in IT/Telecom.
- Most active investor in early stage IT companies in Atlantic Canada.
- Senior Partners include:
  - Bell Aliant
  - Cisco Systems Canada
  - Nortel Networks
- We are actively seeking Research collaborations

The Project

TARA has partnered with a group of companies in a multi year project to analyze the outbound and inbound traffic in Networks of Convenience.

The specific companies and specific objectives of The Project remain confidential at this point.

However, From an analysis perspective we are first interested in understanding the nature of this traffic.

Data sources are real traffic captures from hotels, airports and general hotpots from around the world.

The Project

Networks of convenience are a relatively new and rapidly growing sector of the ISP community.

These are networks that serve a transient population.

The provider is compensated either by fees charged to end users, or by the hosting organization which absorbs the cost as overhead.

The networks may be wired, typically using Ethernet, or wireless (802.11).

Relatively little is known about the ways in which these networks are used.

TARA

The Project

We believe that Networks of Convenience may be used by criminals and / or terrorists in attempts to conceal their activities, identities, or both.

Networks of convenience are the "payphones" of the twenty-first century. Users of these networks take advantage of the implicit anonymity that comes with their use.

We do not know how common other forms of malicious activity may be in these networks

TARA

The Project

Network traffic characterization approaches in the past have relied on availability stable data in an environment of perfect information.

An analyst could have access to static IP and MAC address databases or DHCP lease logs that could be used to collate traffic to specific origins such as identifiable workstation/user combinations, servers or other network attached devices.

In this environment, normal-versus-anomalous behaviour models could be used to profile network and user behaviour to detect misuses or anomalous behaviour such as masquerade attack or worm propagation.

TARA

Data Gathering

Since the sources tend to be NAT'ed, we use network taps on the interfaces inside of the edge router. Currently capturing inbound and outbound data separately.

Prior to analysis, full packet captures are first converted to primitive flows.

Our research is focused on flow level analysis but this conversion also helps to allay provider's concerns for their customer's privacy. (i.e. we don't look at your data only the packet header)

TARA

Observations During Conversion

100 Internal IPs Monitored for 1 month.

Of all Packets Read:

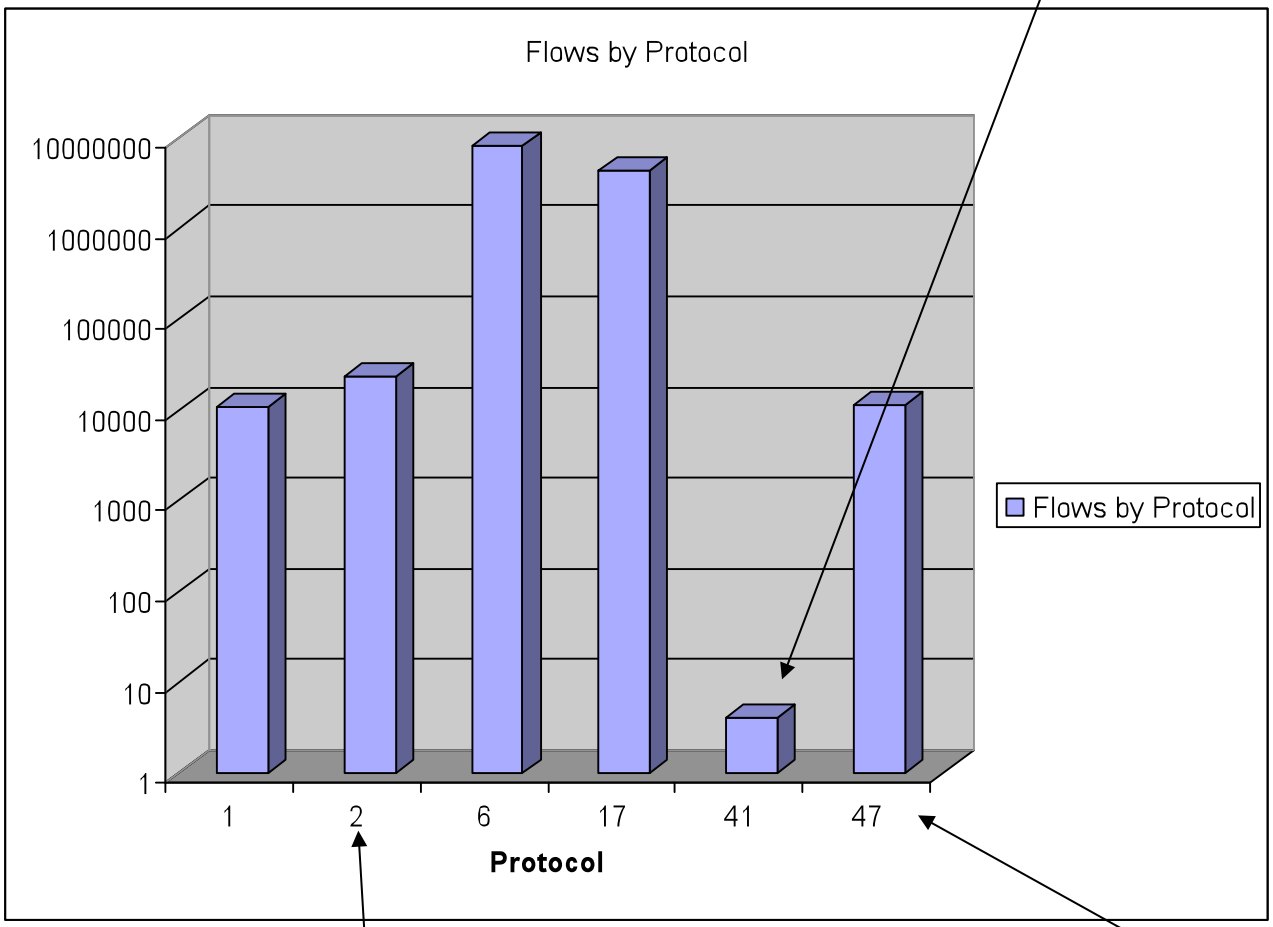
- Not IPV4: 1.7%
- Fragmented: 0.06%
- Too Short: 0.0%
- Incomplete (No Ports and or Flags): 0.0%

Overall, traffic is characterised by its non-uniformity.

TCP=65%
UDP=34%

Protocol Flows were a Little Unusual

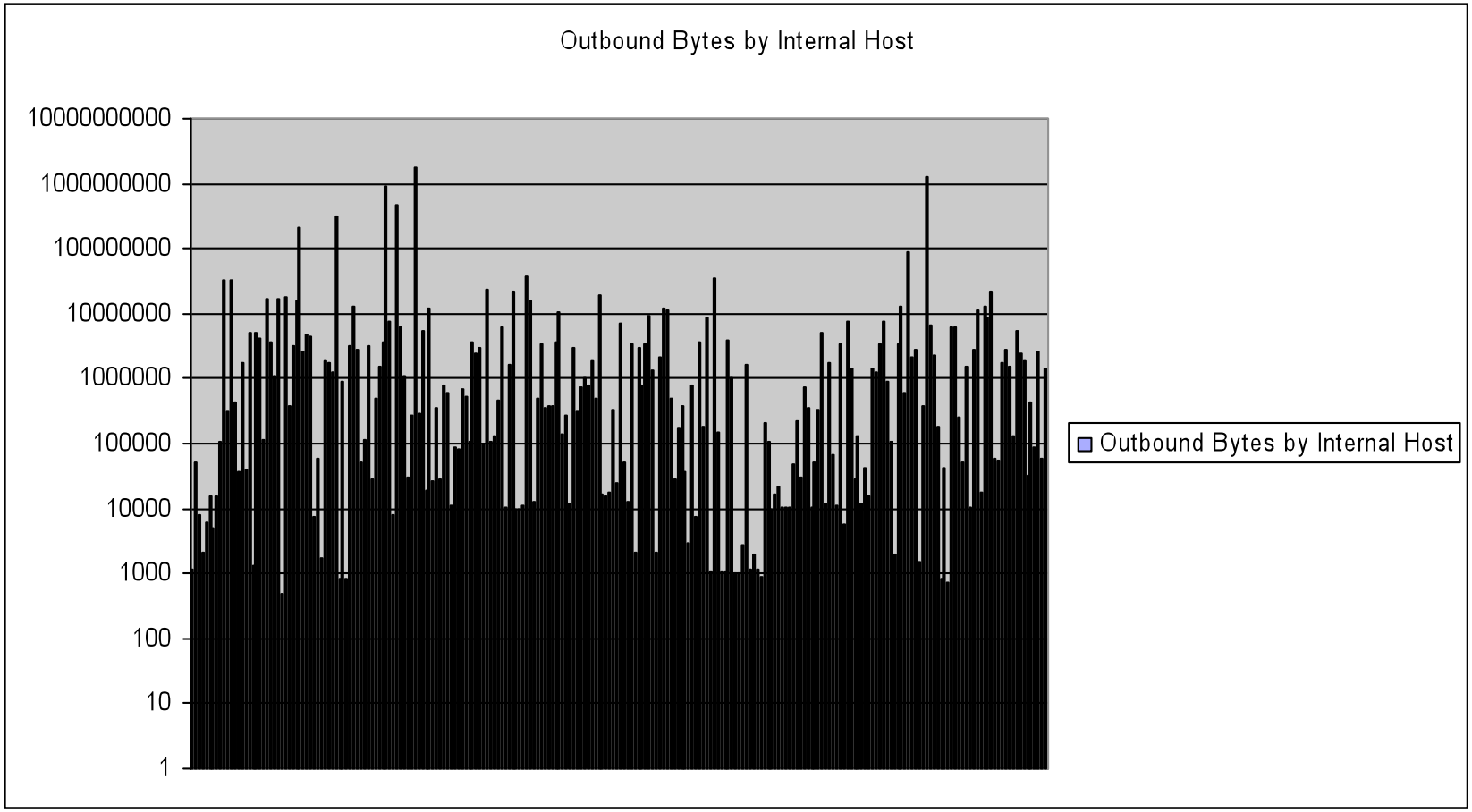
IPv6 Encapsulation
At 0.00003%



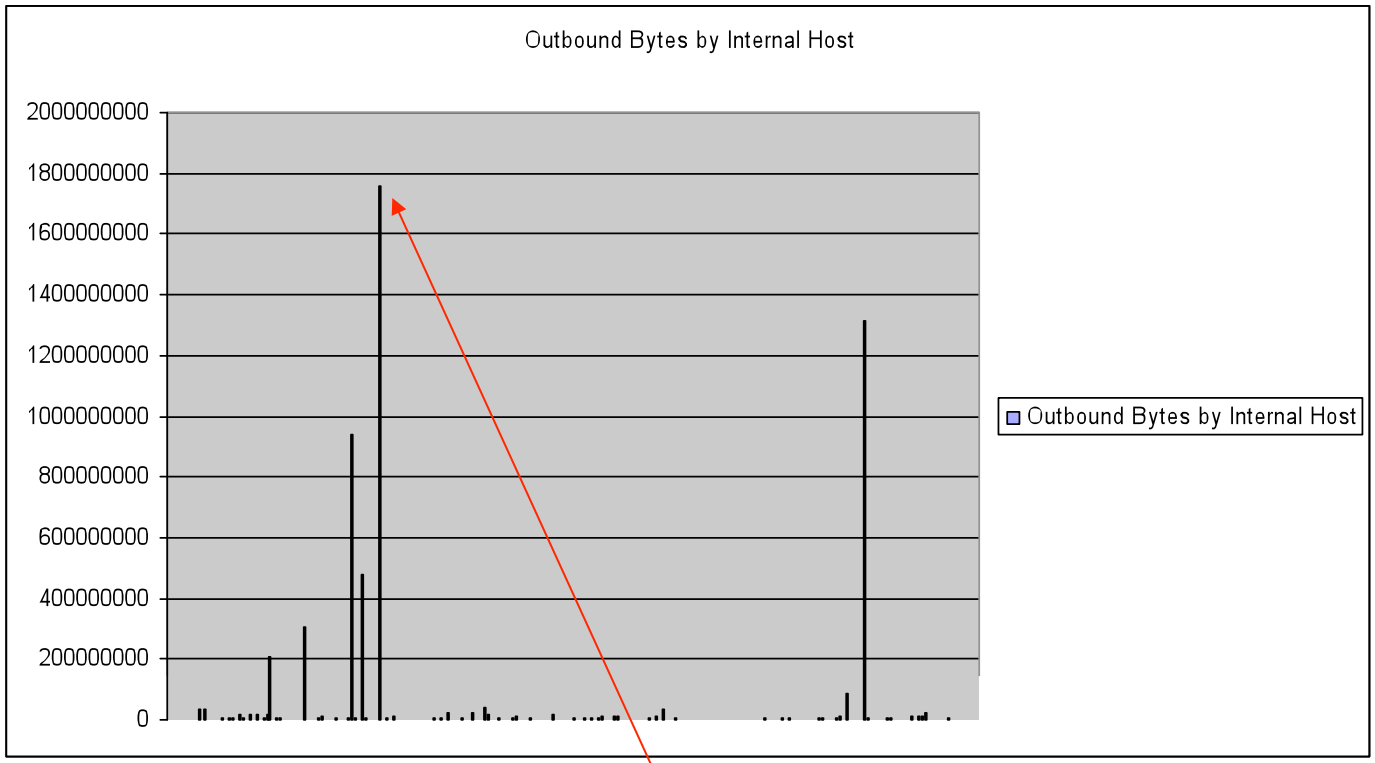
Multicast Host management
At 0.18%

VPN's smaller than I expected at 0.09 %

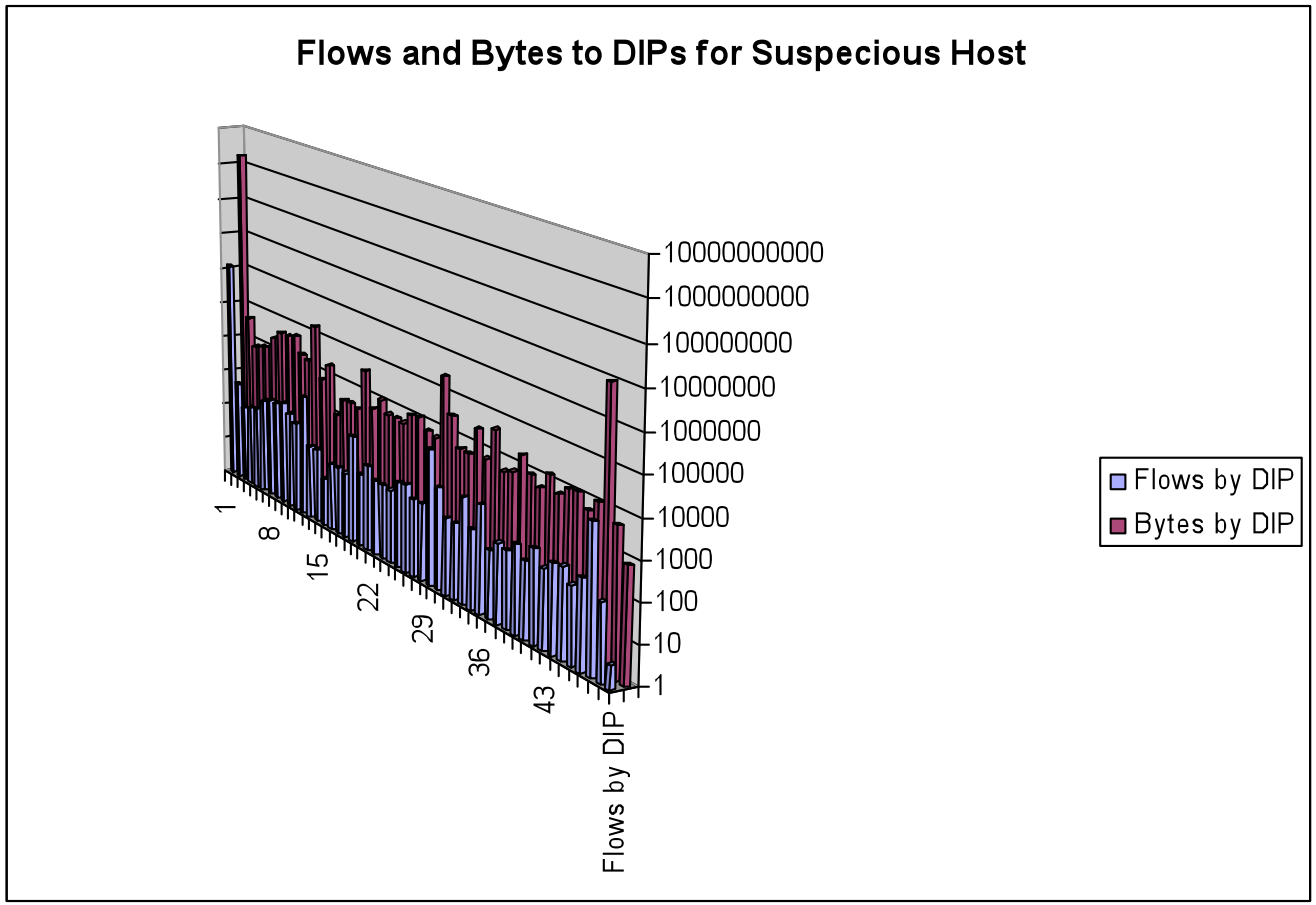
Outbound Bytes by Host Show Large Variations

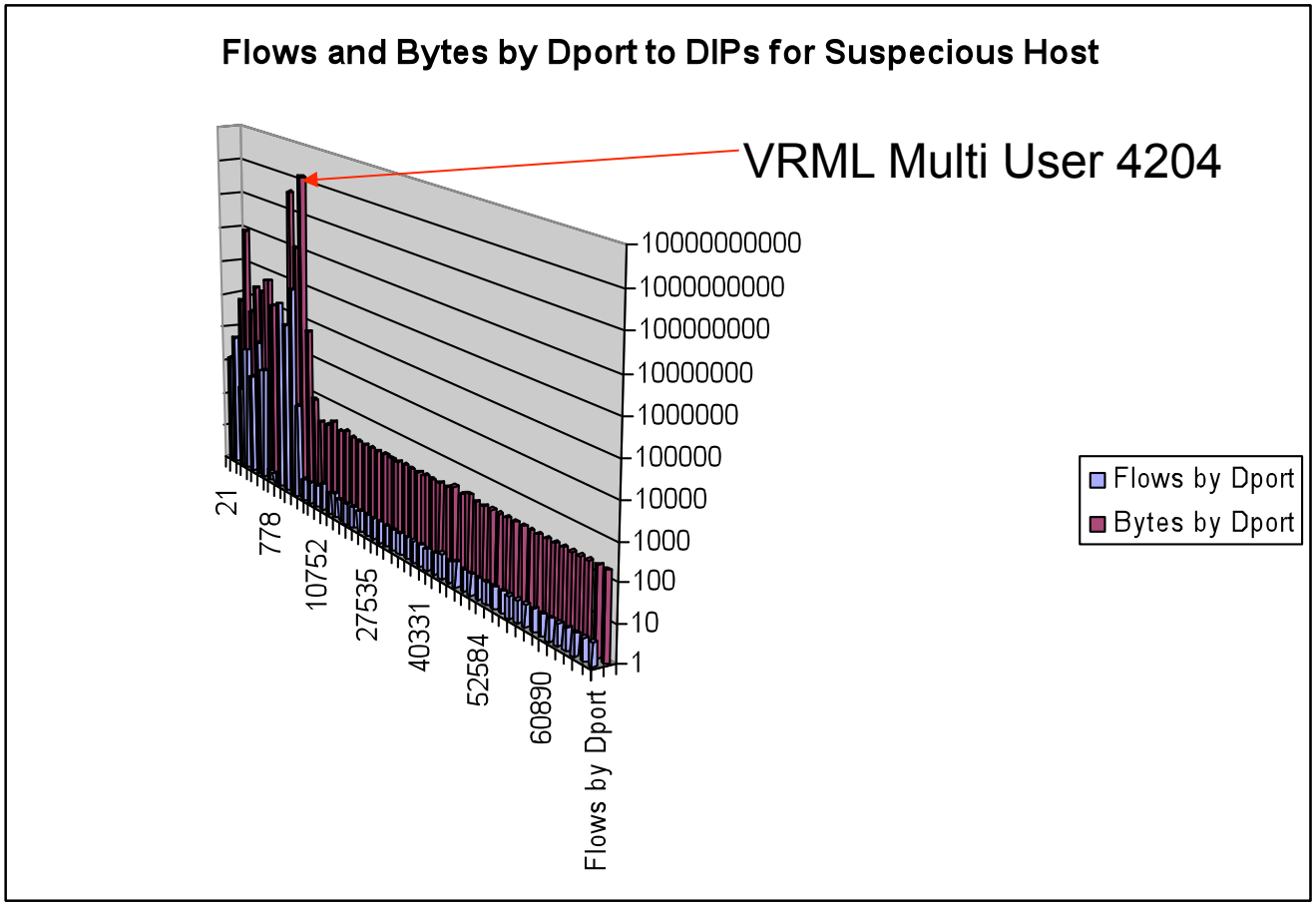


Obvious in a Linear Scale

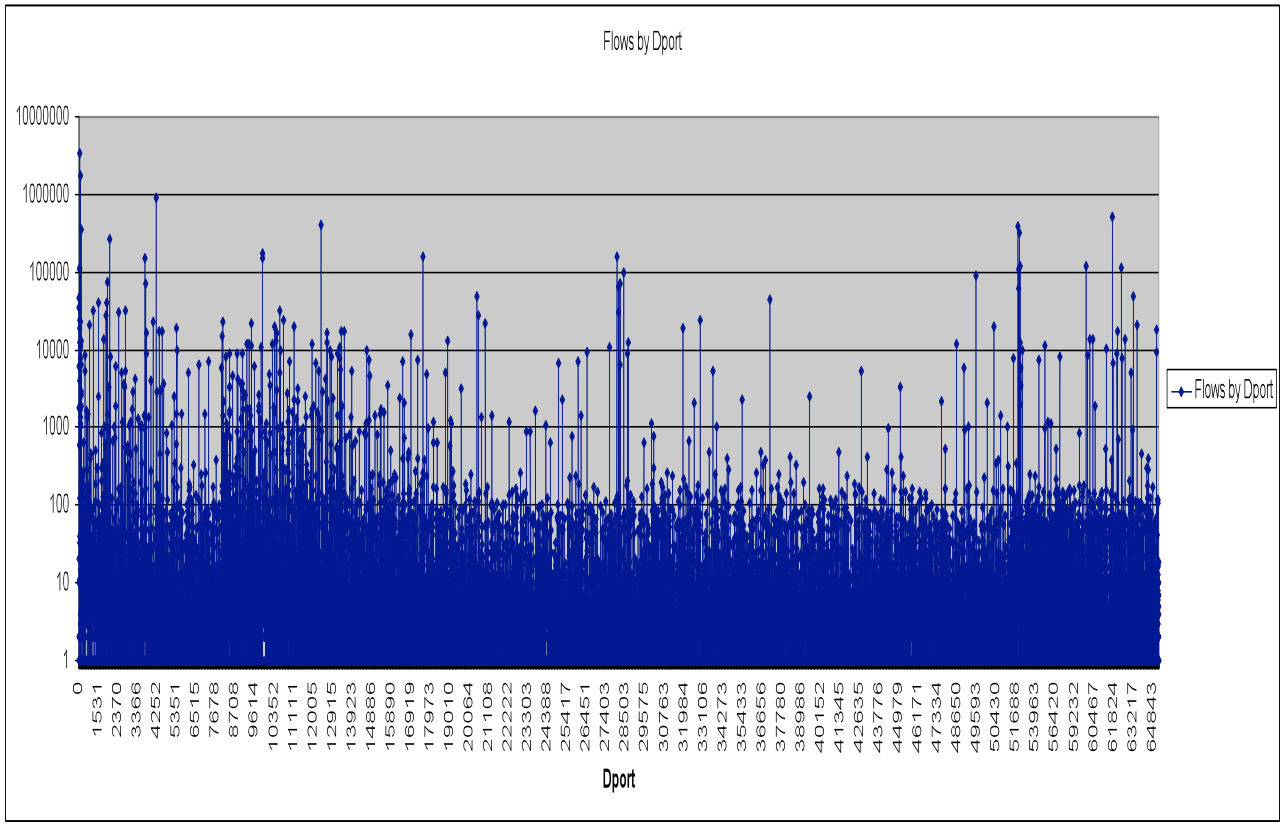


Lets take a closer look at this guy


Flows and Bytes to DIPs for Suspecious Host
10000000000
1000000000
100000000
10000000
1000000
100000
10000
1000
100
10
1
1
8
15
22
29
36
43
Flows by DIP
Flows by DIP
Bytes by DIP


Flows and Bytes by Dport to DIPs for Suspecious Host
VRML Multi User 4204
10000000000
1000000000
100000000
10000000
1000000
100000
10000
1000
100
10
1
21
778
10752
27535
40331
52584
60890
Flows by Dport
Flows by Dport
Bytes by Dport

We expected DPorts 80 and 443 to represent most traffic….

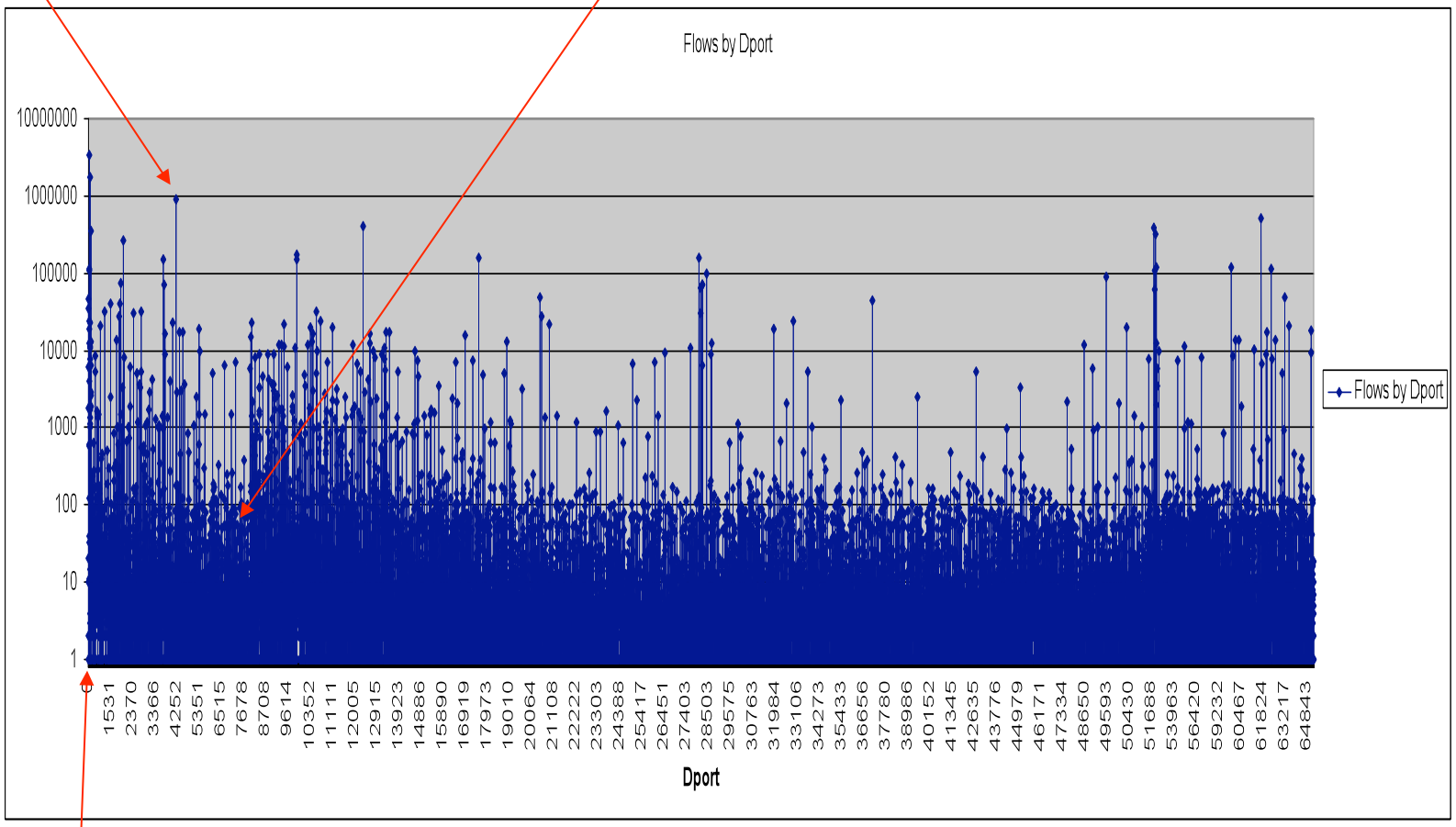


Together they accounted for 41%

Note that the DPort 0 data point is ICMP Traffic

4204 lists as VRML Multi-User almost all from 1 host

One host only 25 flows on BitTorent



Only minute traces of Half Life Gaming

Note that the DPort 0 data point is ICMP Traffic

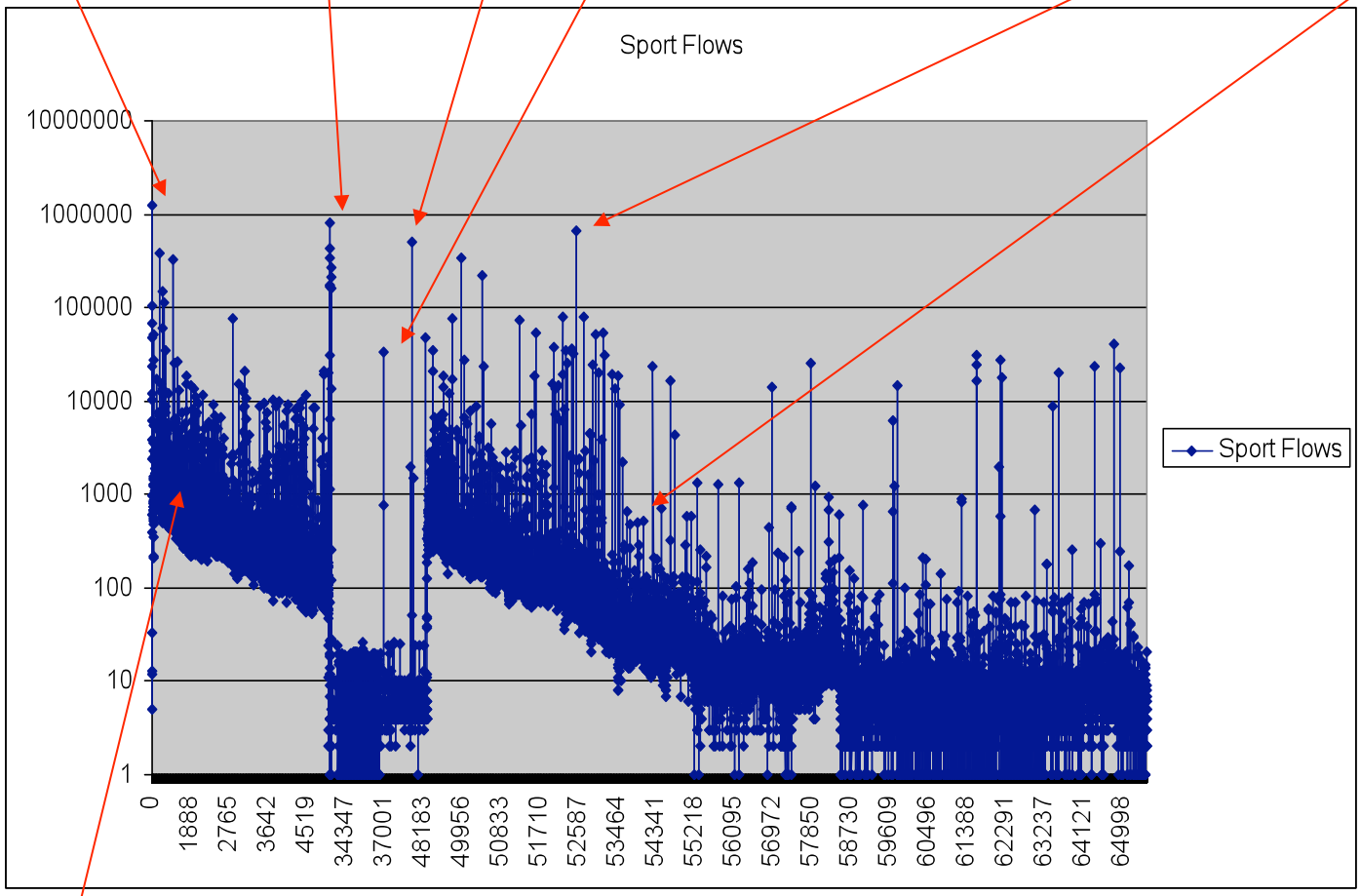
SSH 10% of all flows

13991 and 44849

36459

52523

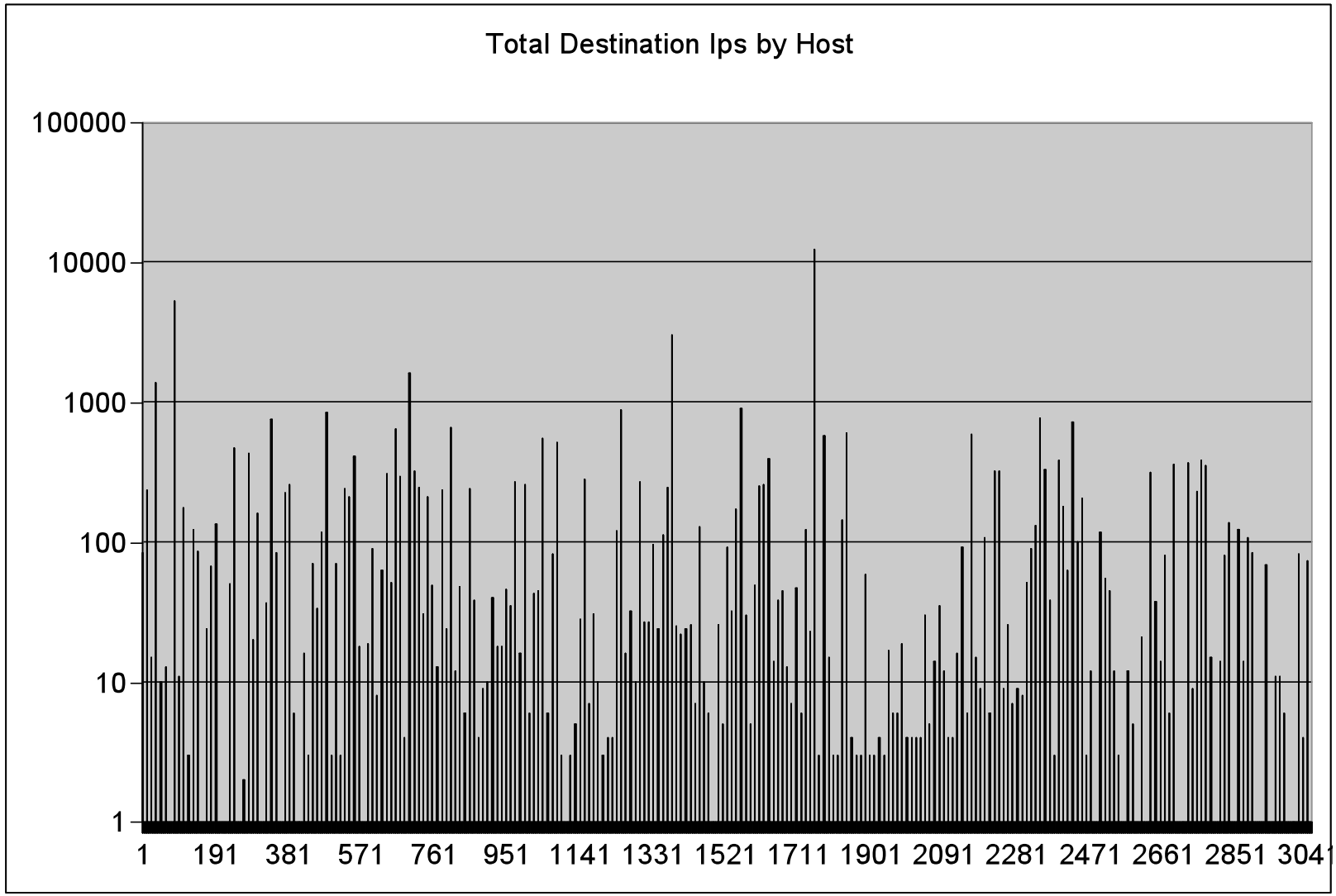
Mac Skype



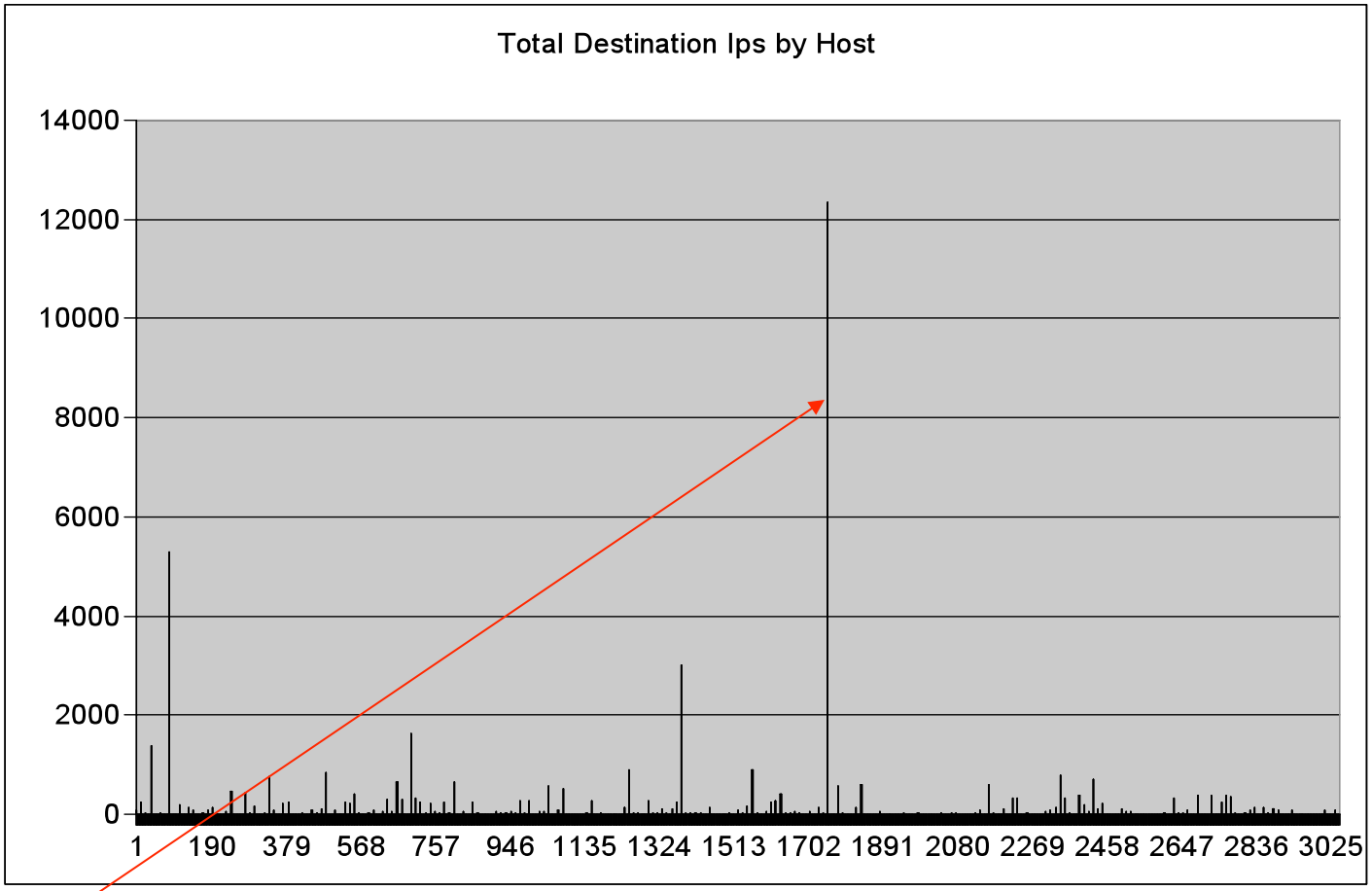
Fasttrack 50 hosts 1700 flows

No 6667 listening?

Number of Destination by Host Shows Substantial Spikes

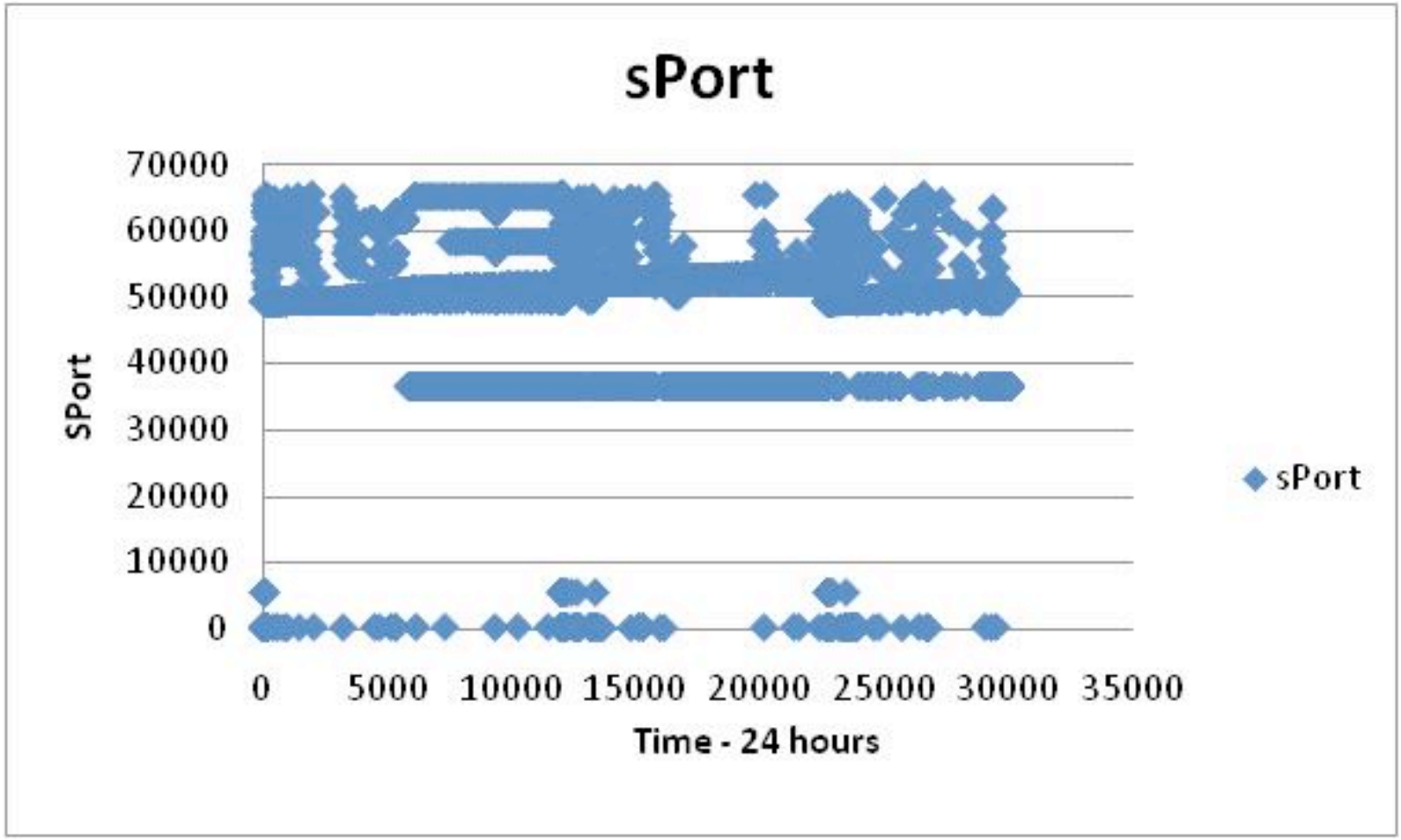


Linear Scale

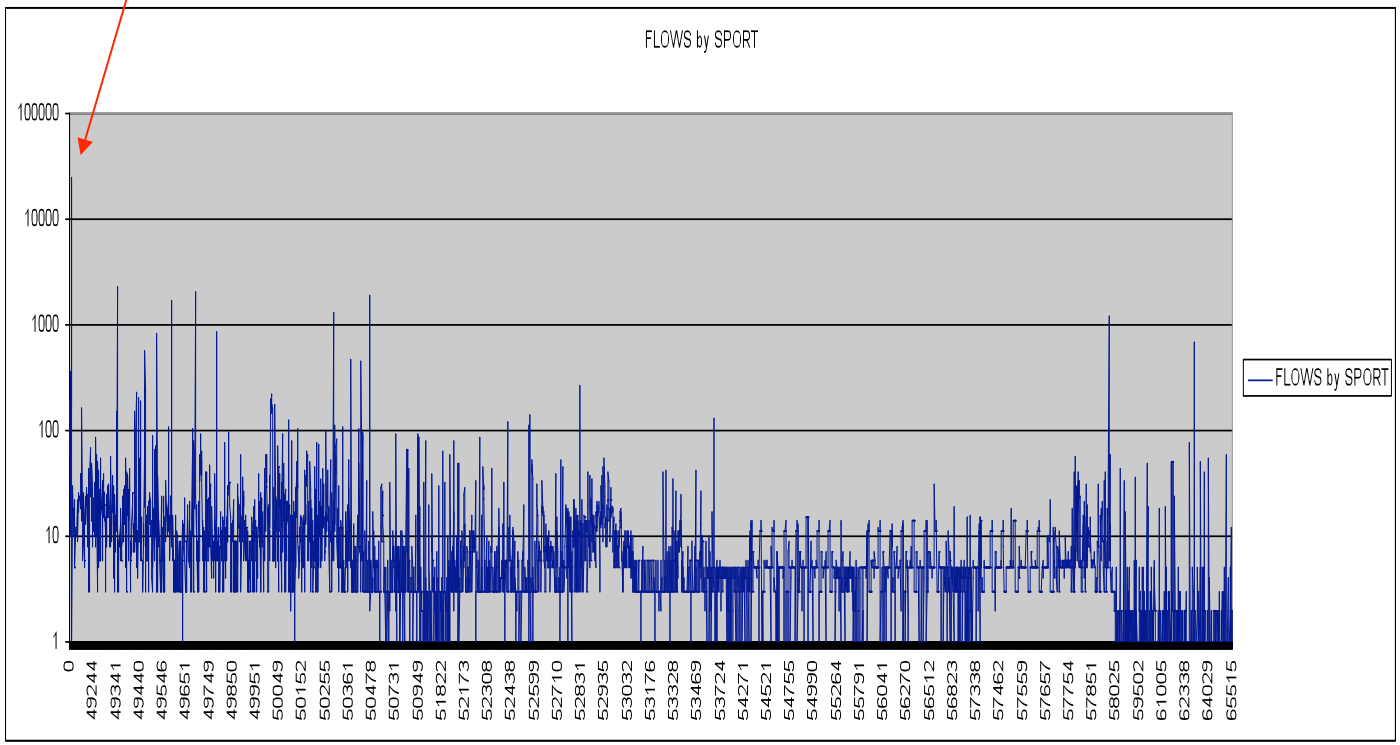


Suspicious Host accessed sequential ranges through multiple /16`s

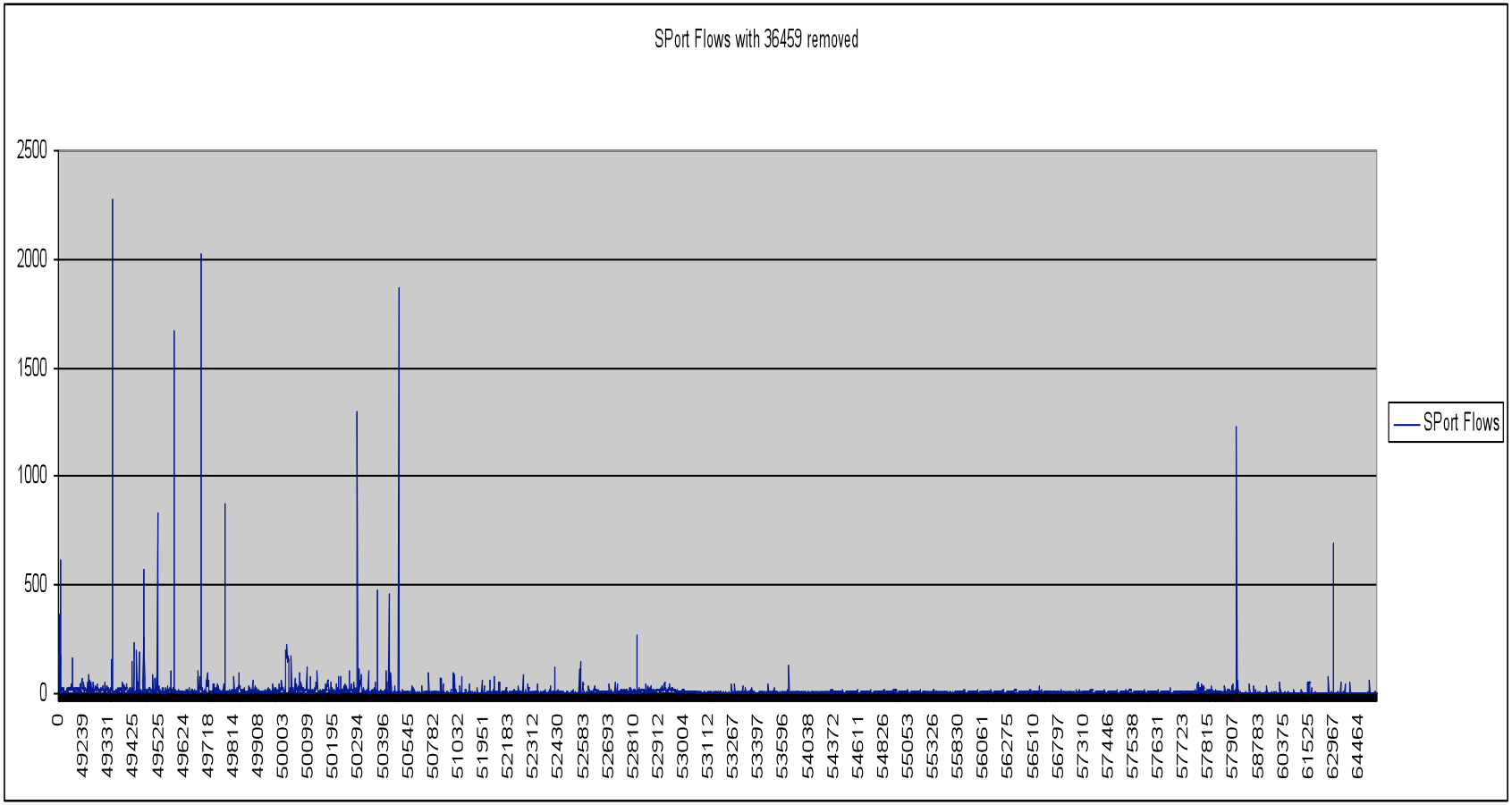
Lets look a little closer at his activity


sPort
70000
60000
50000
40000
30000
20000
10000
0
SPort
0
5000
10000
15000
20000
25000
30000
35000
Time - 24 hours
sPort

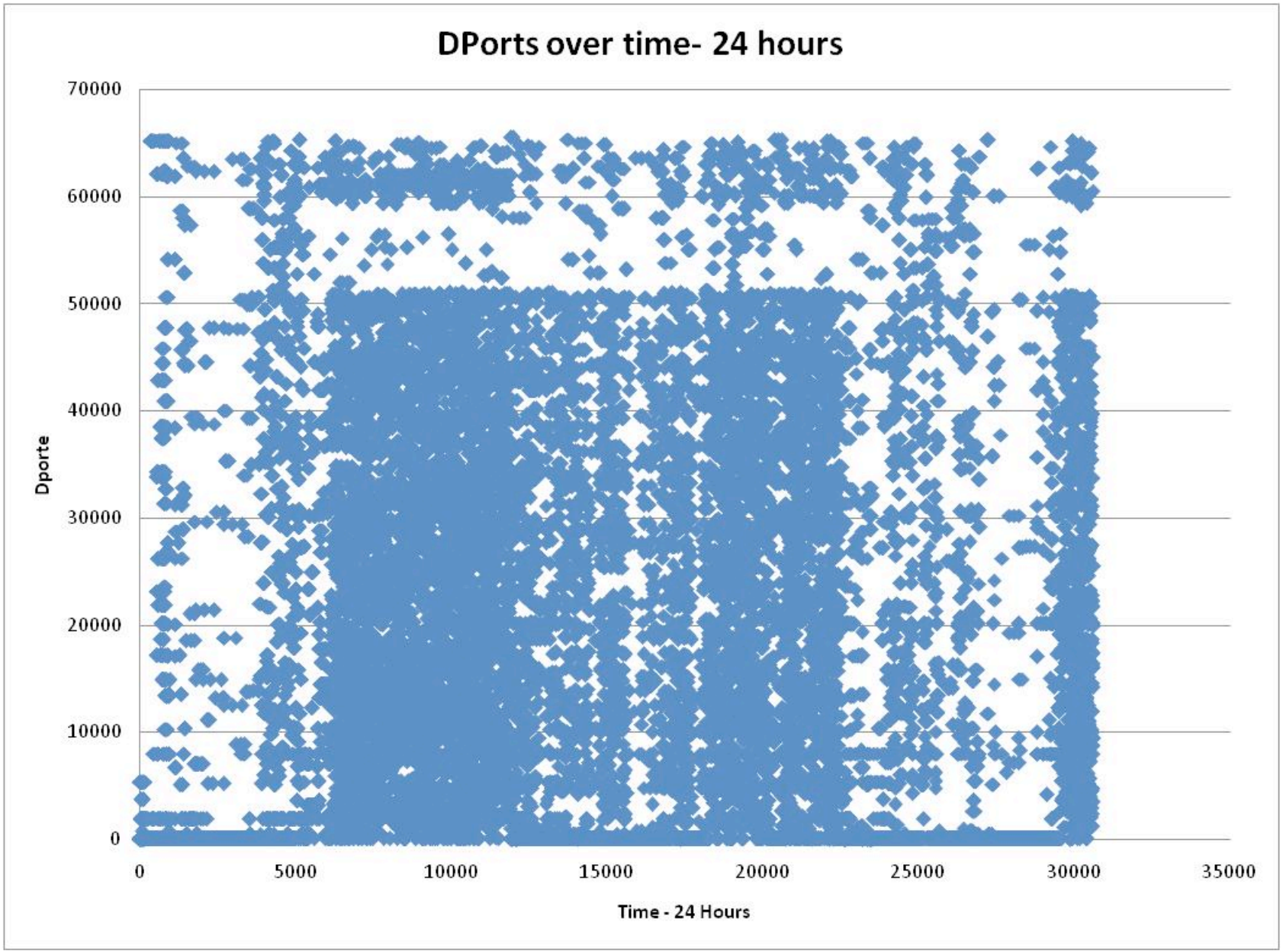
36459 Dominates the Sports…



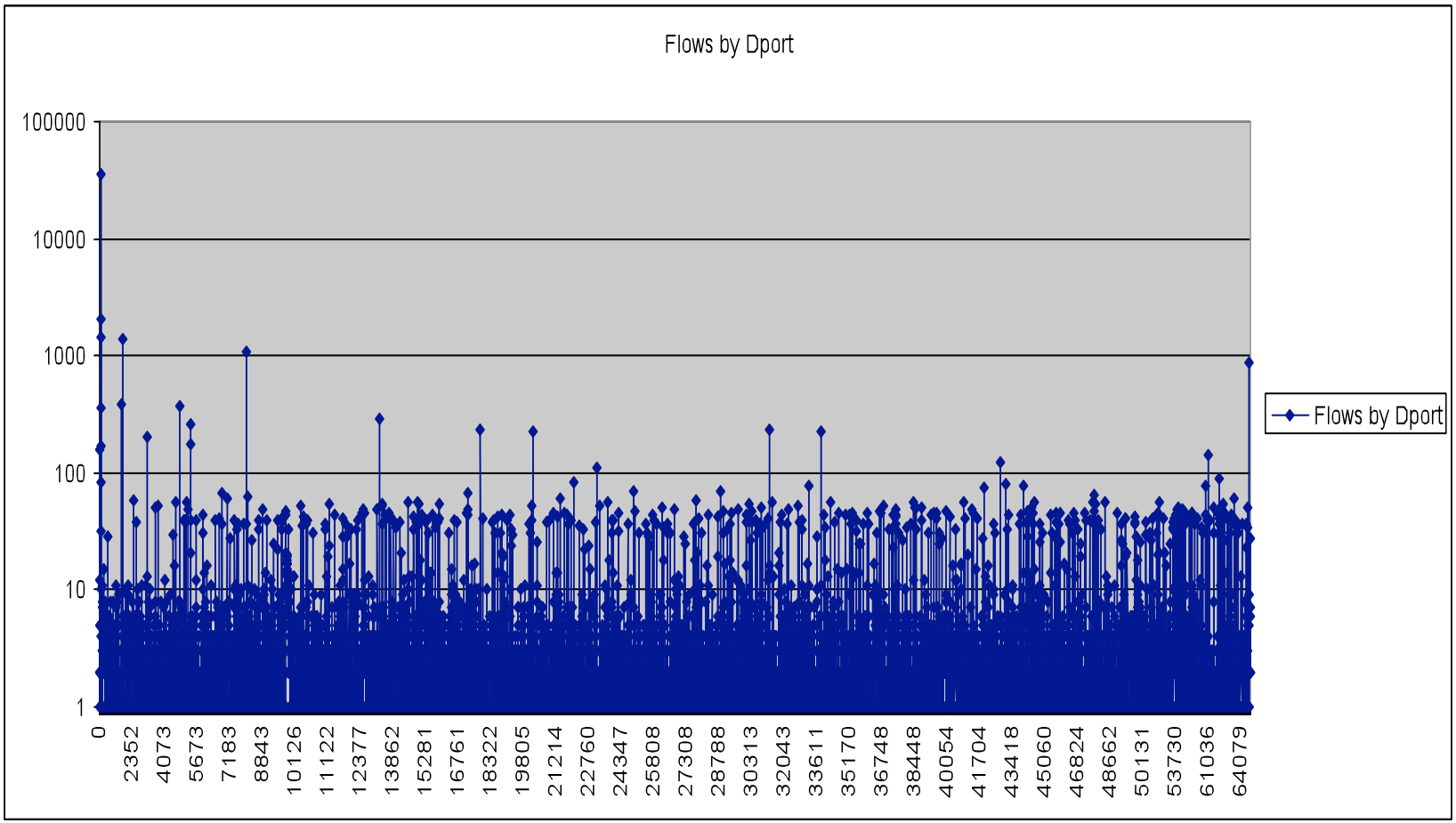
Removing 36459 in a linear scale we get



Sport use is near sequential above 49153


DPorts over time- 24 hours
70000
60000
50000
40000
30000
20000
10000
0
Dporte
0
5000
10000
15000
20000
25000
30000
35000
Time - 24 Hours

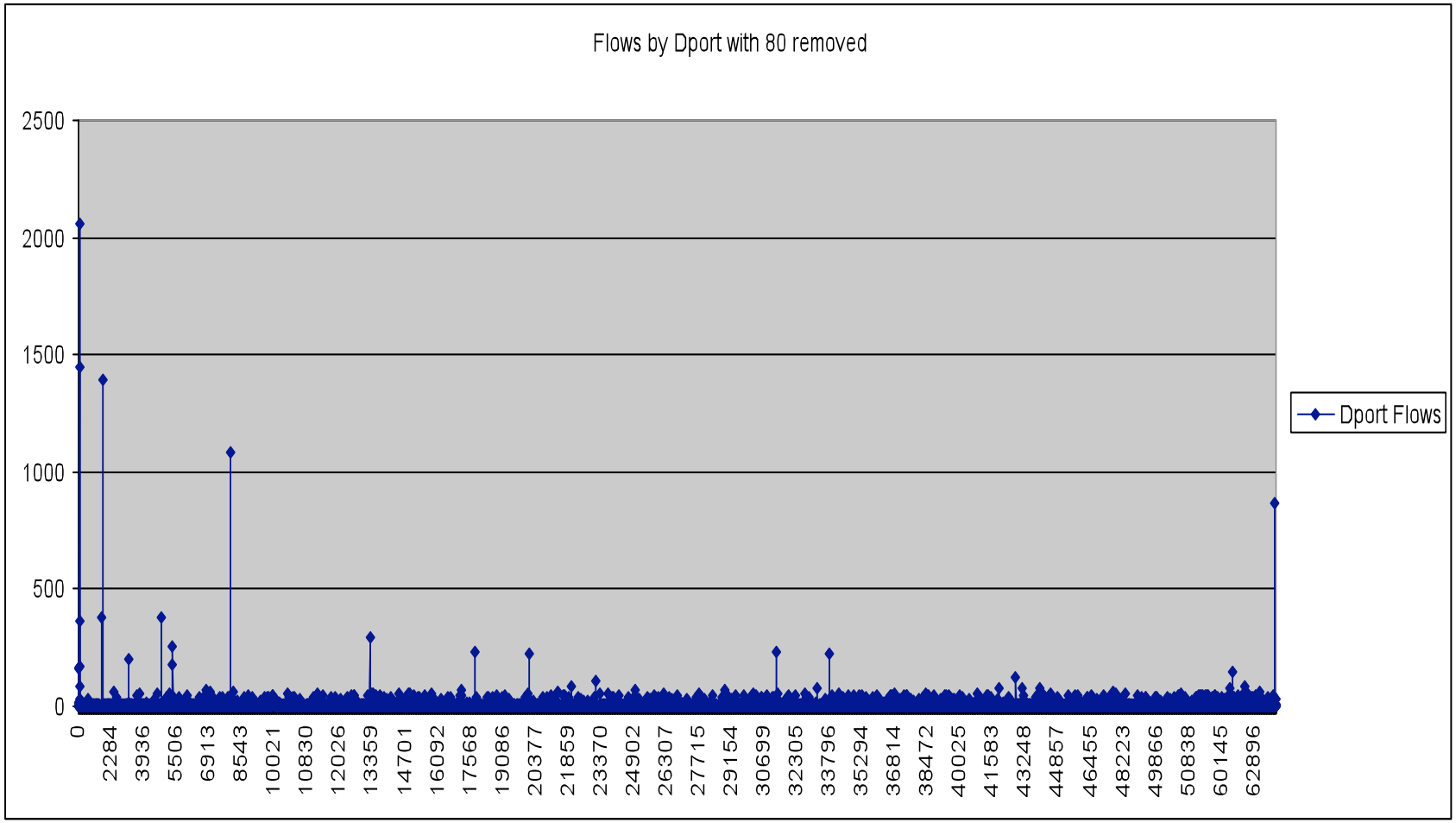
His DPort values in Log Scale

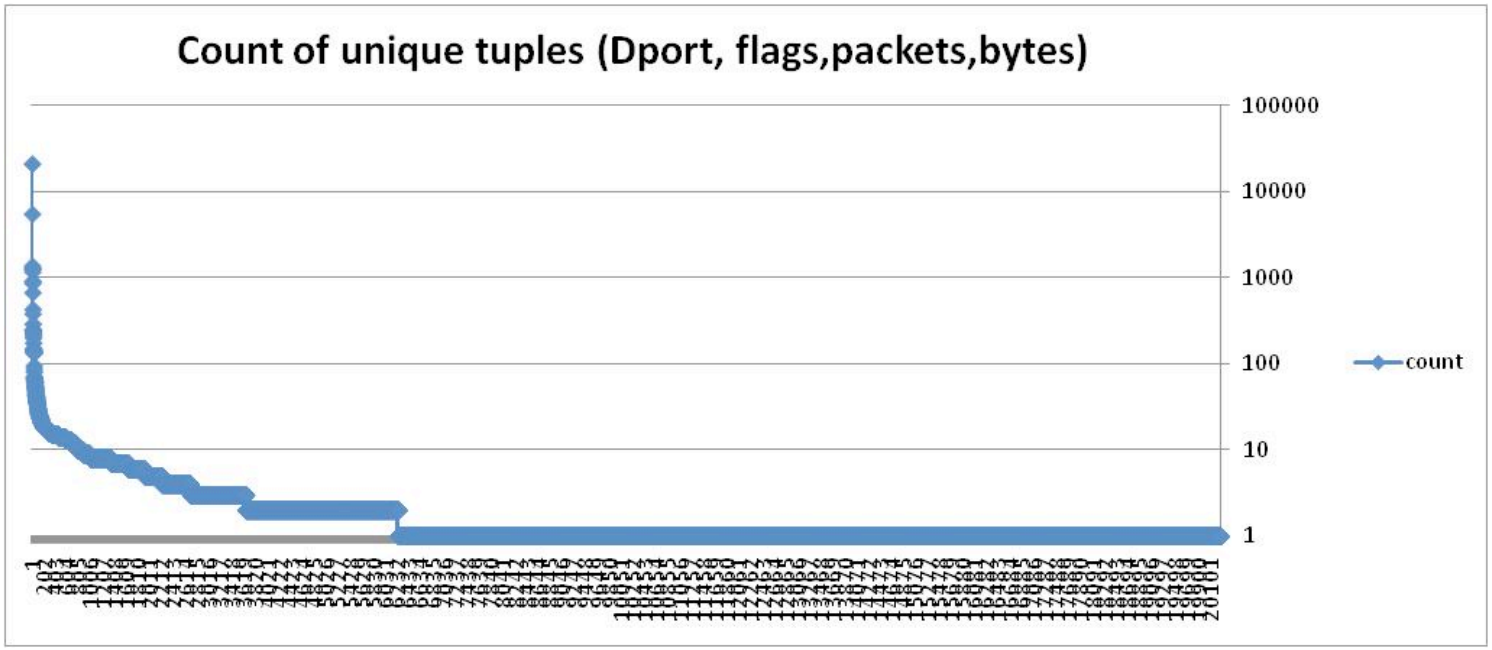


Not quite sequential but most ports above 1024 are accessed.

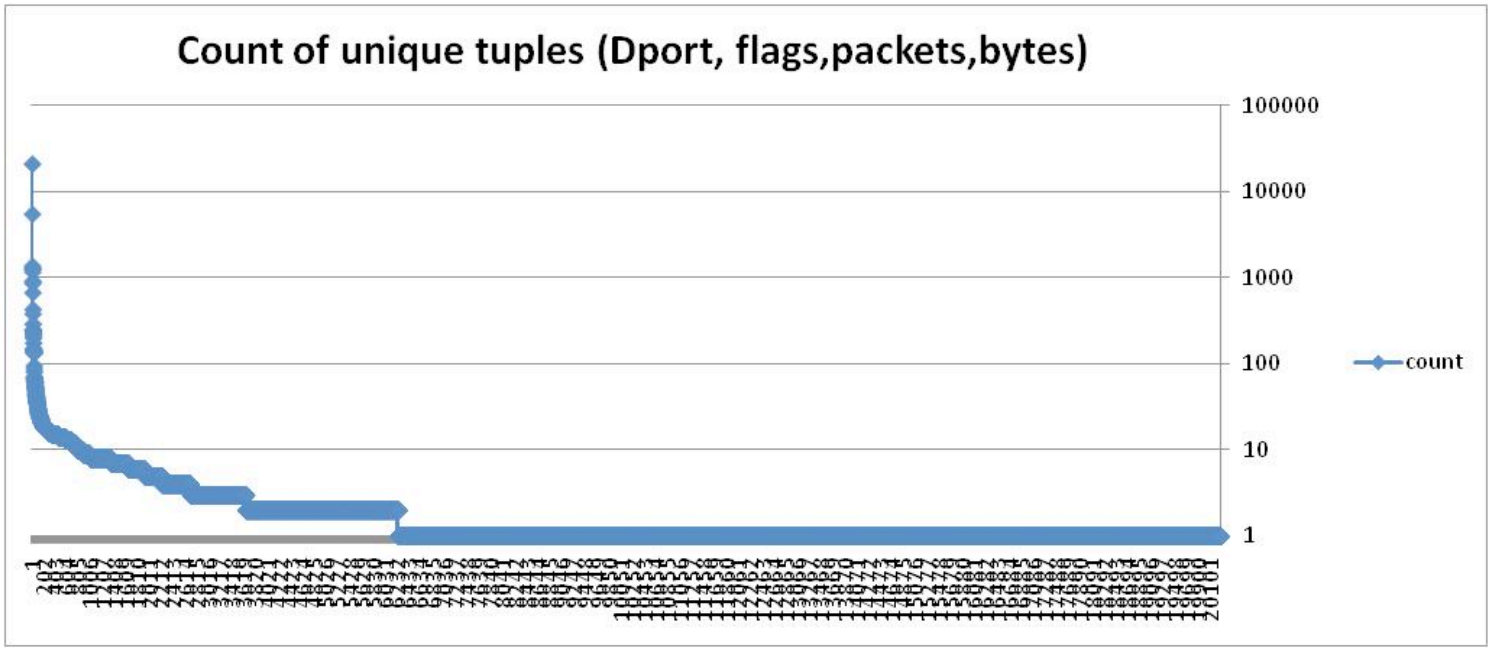
Average is less than 10 flows (packets) per Dport

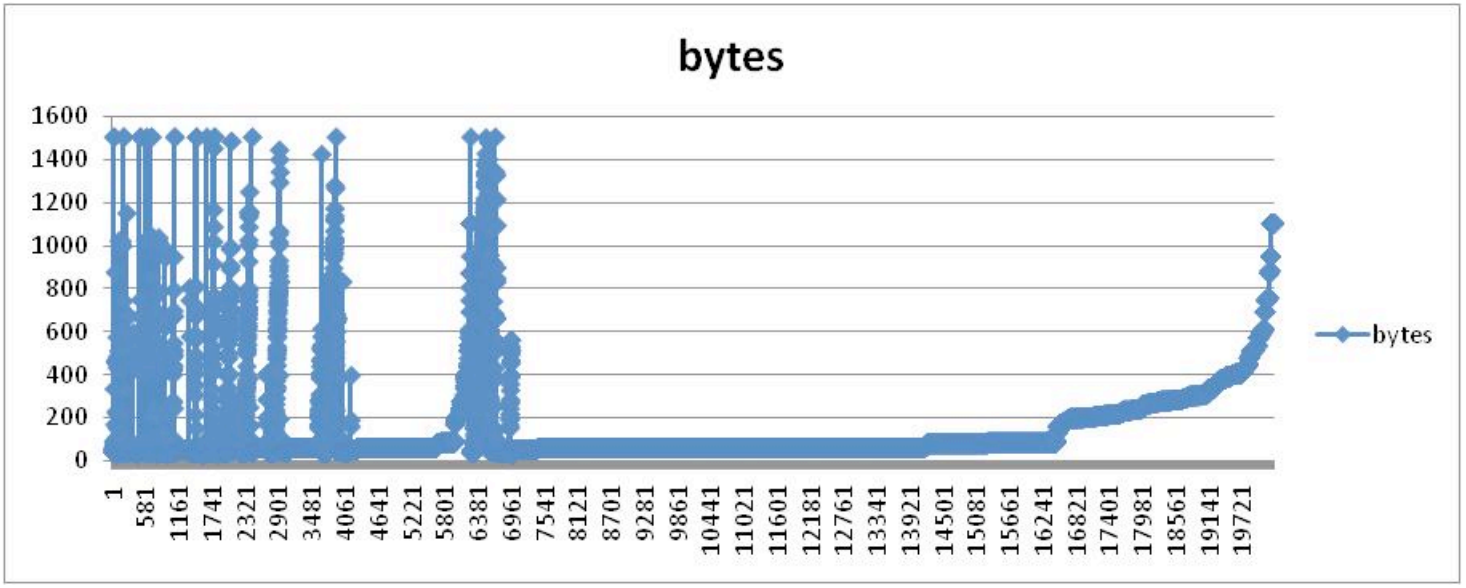
Linear version Dports with Port 80 removed

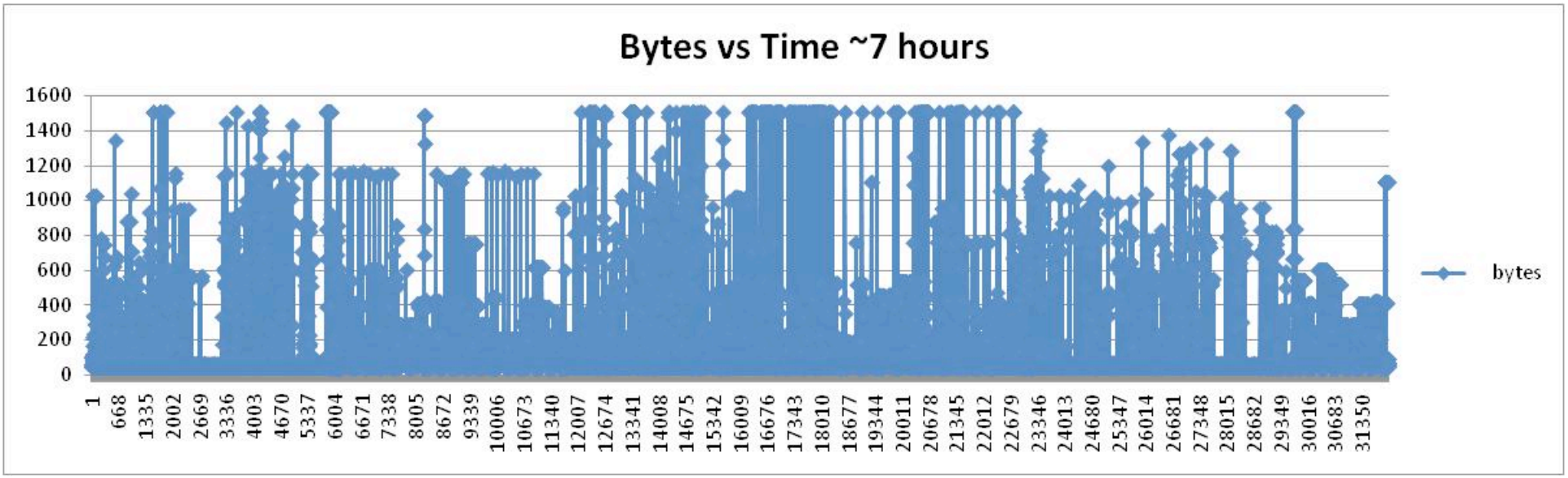




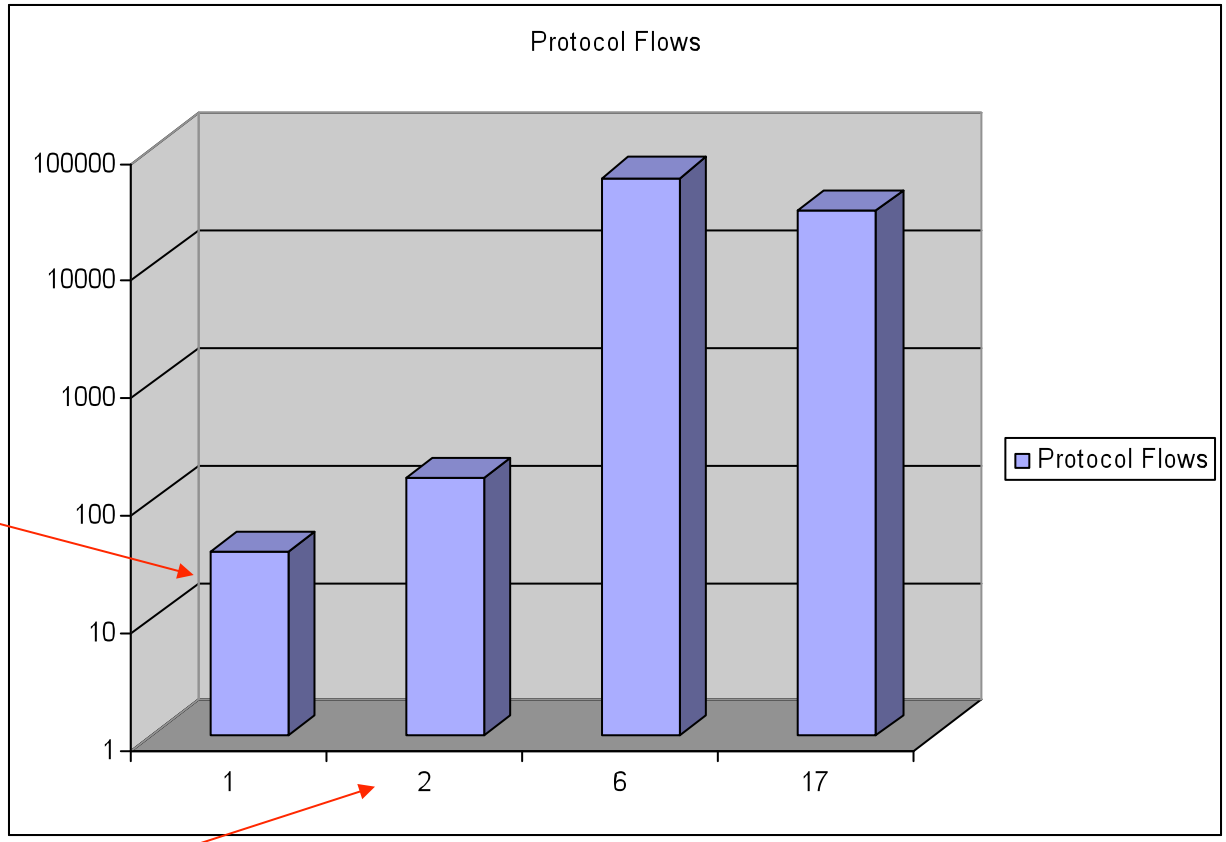
count	pro	dPort	flags	packets	bytes
21005	6	80	A	1	40
5502	6	80	A	1	52
1373	6	443	A	1	40
1280	6	80	RA	1	40
1261	17	1900		1	61
1173	6	80	S	1	52
914	6	80	S	1	48
866	6	65209	R	1	40
673	6	80	FA	1	40
430	6	80	A	1	60


Count of unique tuples (Dport, flags,packets,bytes)
100000
10000
1000
100
10
1
count


bytes
1600
1400
1200
1000
800
600
400
200
0
bytes


Bytes vs Time ~7 hours
1600
1400
1200
1000
800
600
400
200
0
1
668
1335
2002
2669
3336
4003
4670
5337
6004
6671
7338
8005
8672
9339
10006
10673
11340
12007
12674
13341
14008
14675
15342
16009
16676
17343
18010
18677
19344
20011
20678
21345
22012
22679
23346
24013
24680
25347
26014
26681
27348
28015
28682
29349
30016
30683
31350
bytes

Protocol Distribution for Suspicious Host



ICMP ratio (0.08%)
Is double
the aggregate
value (0.04%)

Multi-Cast Host Management 0.18%

65 % TCP
34 % UDP

Would expect these to be more equal for a peer and vastly skewed for a scanner.

Tradition Demands that I ask this Question

Massive Destination IP`s (sequential /16`s)
Massive Source Ports (near sequential)
Massive Destination Ports (near sequential)
Multi-Cast Host Management Protocol
Larger than expected ICMP Ratio
Standard TCP/UDP Ratio

WHO AM I ?

Tradition Demands that I ask this Question

Massive Destination IP`s (sequential /16`s)
Massive Source Ports (near sequential)
Massive Destination Ports (near sequential)
Multi-Cast Host Management Protocol
Larger than expected ICMP

WHO AM I ?

However, unlike previous years…..

I HAVE NO IDEA…….

Summary

Some obvious challenges

How to tell when host changes?

- Will test user-host profiler presented at flocon 2006.
- until this is nailed down – assumptions are more like ``let`s pretend``.

Some Intriguing Opportunities

- Oops – I`m not allowed to talk about those yet.

Thank You

I am seeking help and would welcome any private feedback, discussions or ideas you might have.

If you had access to this data – what would you do?