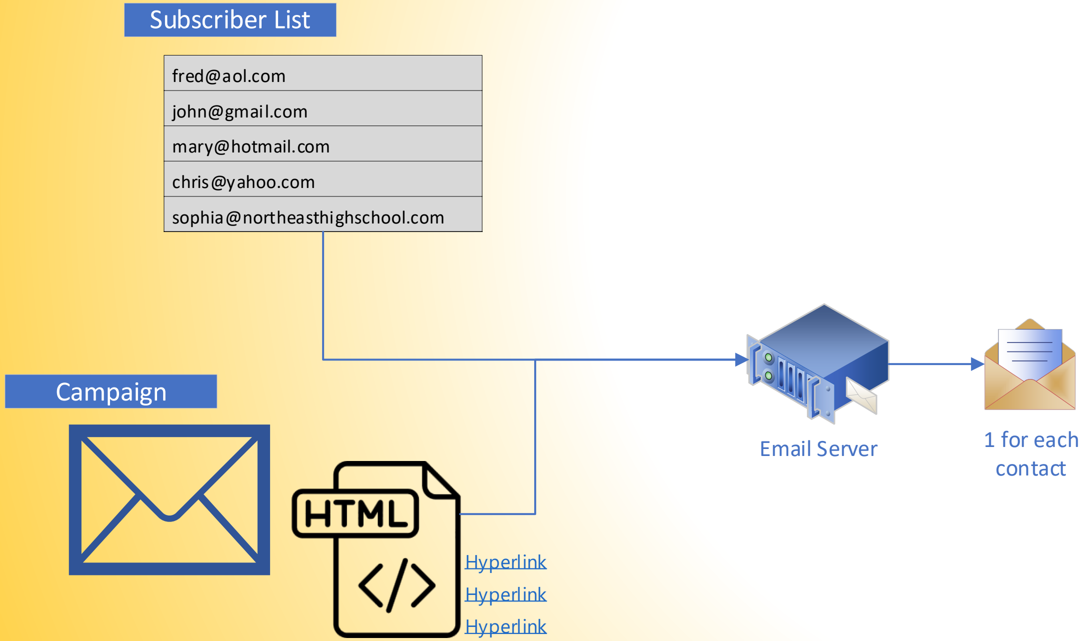
Email Tracking BOT Detection

Background on Email Tracking

- Sending email campaigns can consume a lot of time and money
- But companies that do campaigns successfully can be rewarded with increased business
- So knowing the effectiveness of your email campaigns is a critical step
- There are 2 main ways to get feedback on your email sends
    - Add a tracking pixel to the email that lets us know when the email was opened
    - Change all the links in the email to tracking links
        - When the contact clicks on a link, the request comes to a tracking service and a redirect to the correct URL is returned
        - This click request inform us of when every link in the email was clicked

Campaigns and Subscribe Lists

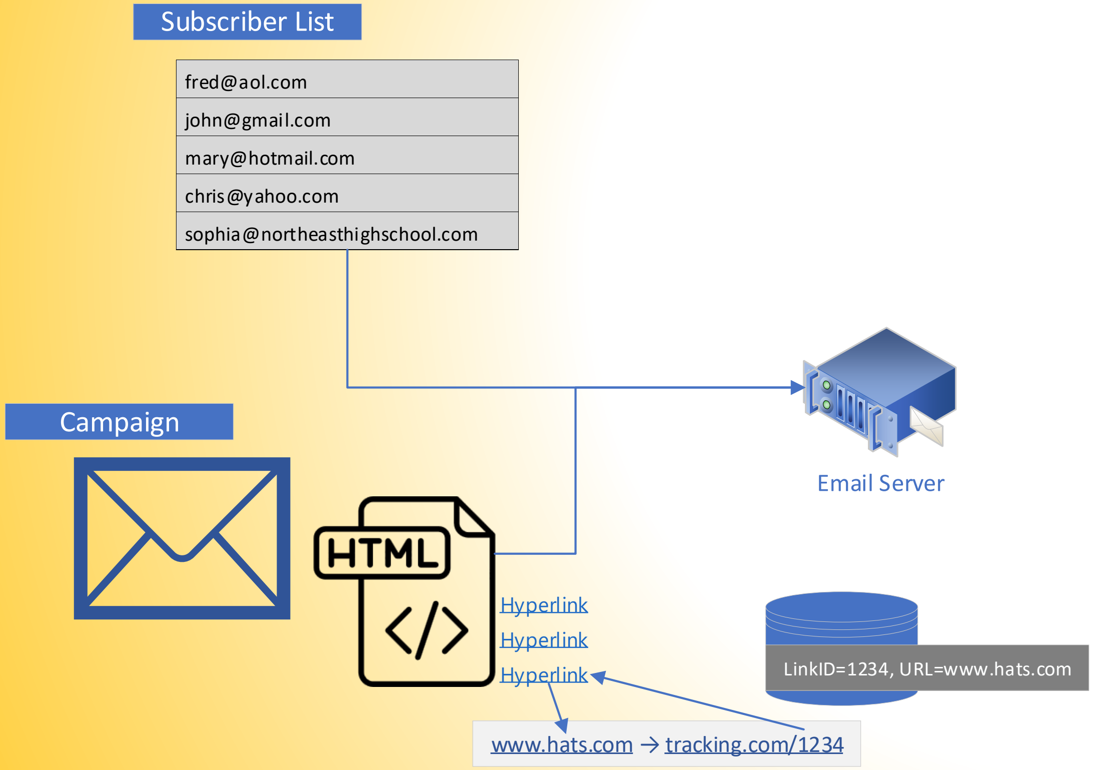


Subscriber List

fred@aol.com
john@gmail.com
mary@hotmail.com
chris@yahoo.com
sophia@northeasthighschool.com

- A campaign is an email sent to the addresses in the subscribe list
- For the most part, every contact receives the same message
- Each campaign can have one or more links in the message

Tracking Links



Subscriber List

fred@aol.com
john@gmail.com
mary@hotmail.com
chris@yahoo.com
sophia@northeasthighschool.com

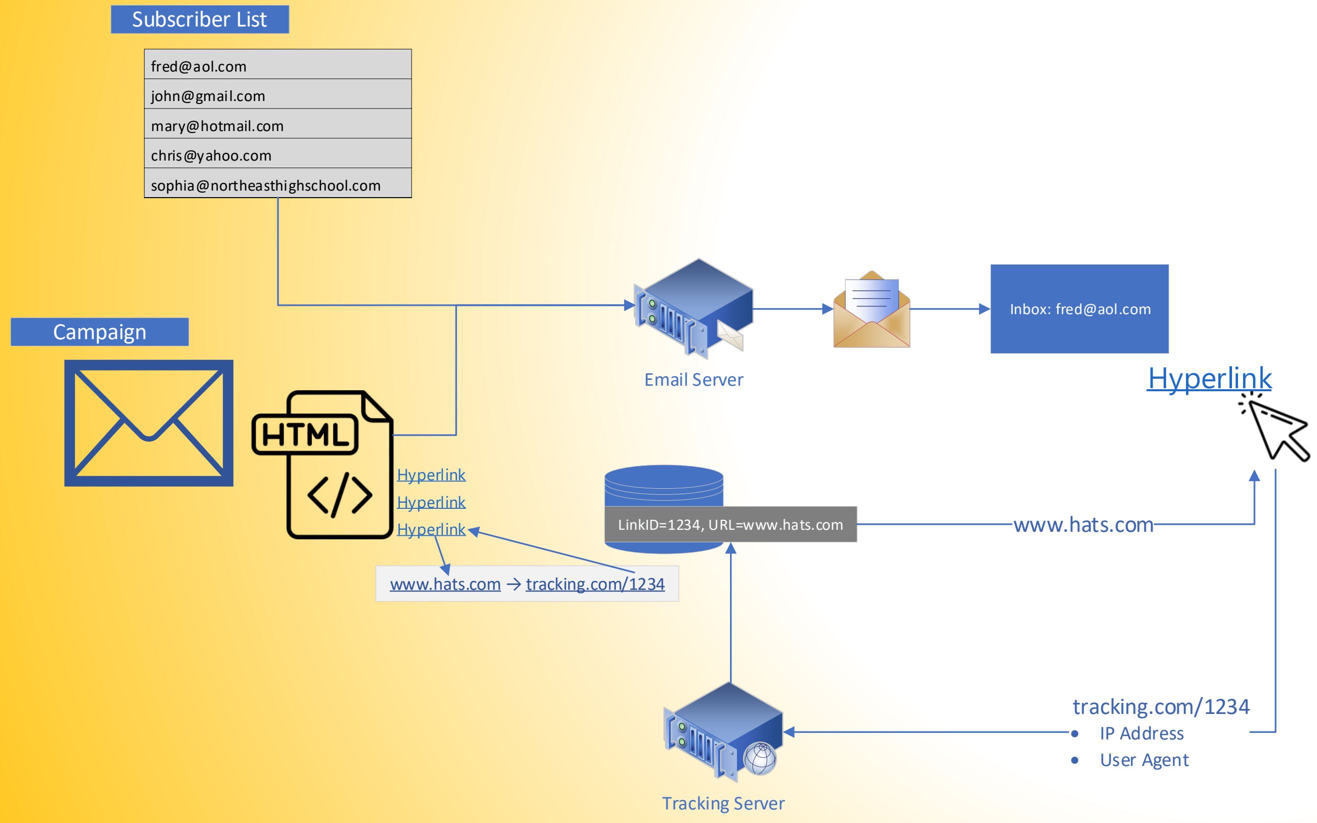
Campaign

LinkID=1234, URL=www.hats.com

www.hats.com → tracking.com/1234

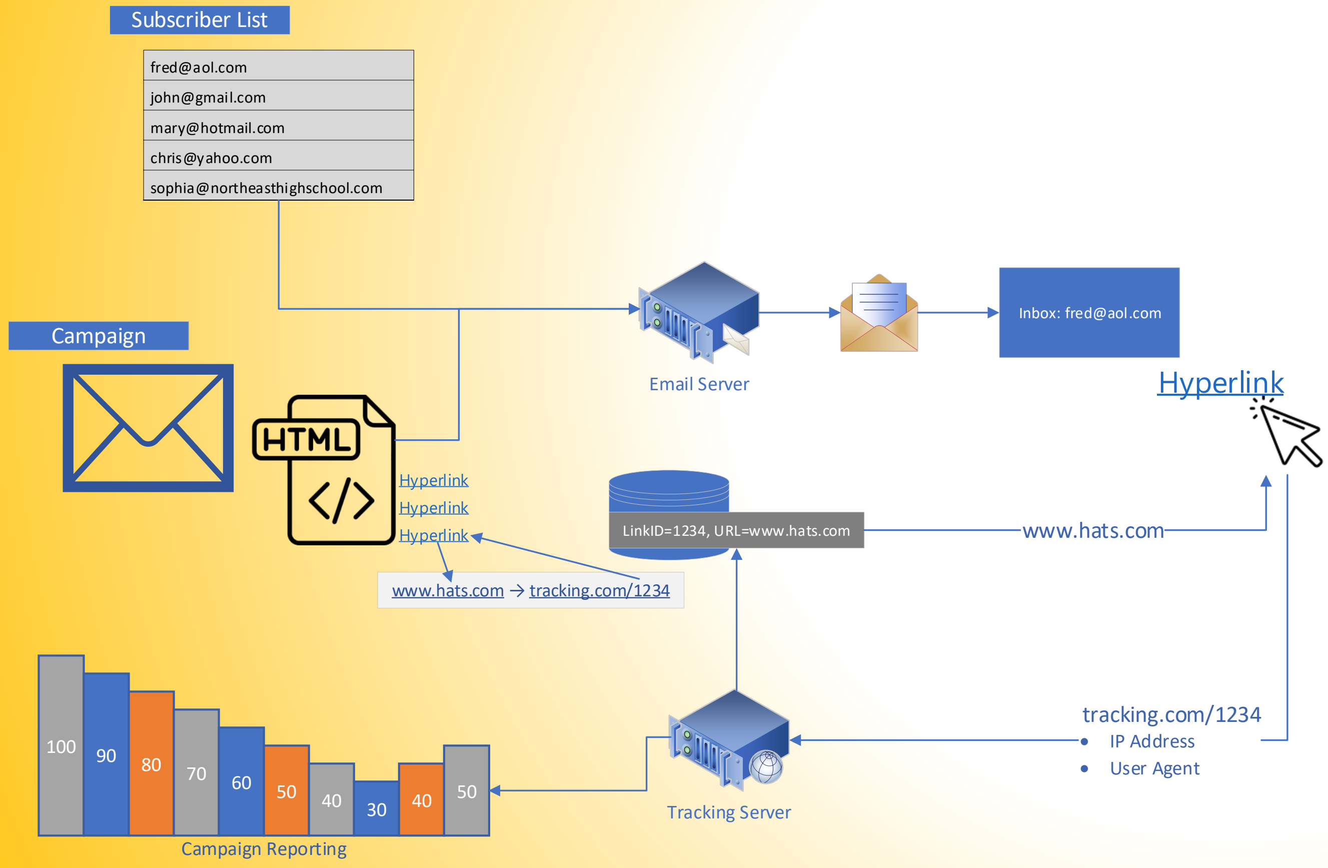
- When the message is generated for each contact, the URL in the links gets replaced with a tracking URL
- The original URL is stored in a database with a unique ID
- The tracking link knows the link ID stored in the database so the original URL is know
- Each tracking link also contains an ID to shows us which contact the click belongs to

Tracking Links



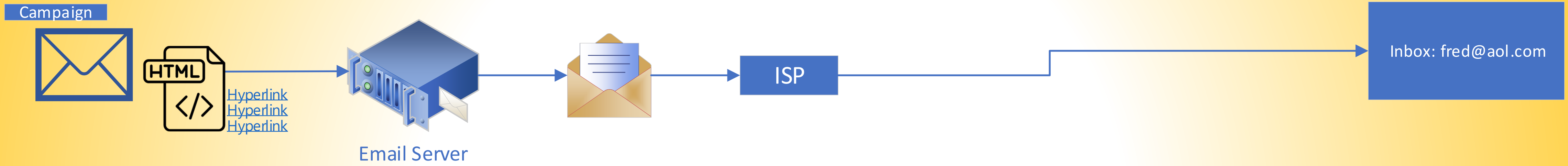
- When the contact clicks on a link, they are sent to the tracking server
- The tracking server looks up the original URL form the database
- A redirect to the original URL is sent back to the contact
- This automatically sends the contact to the proper URL

Reporting Clicks



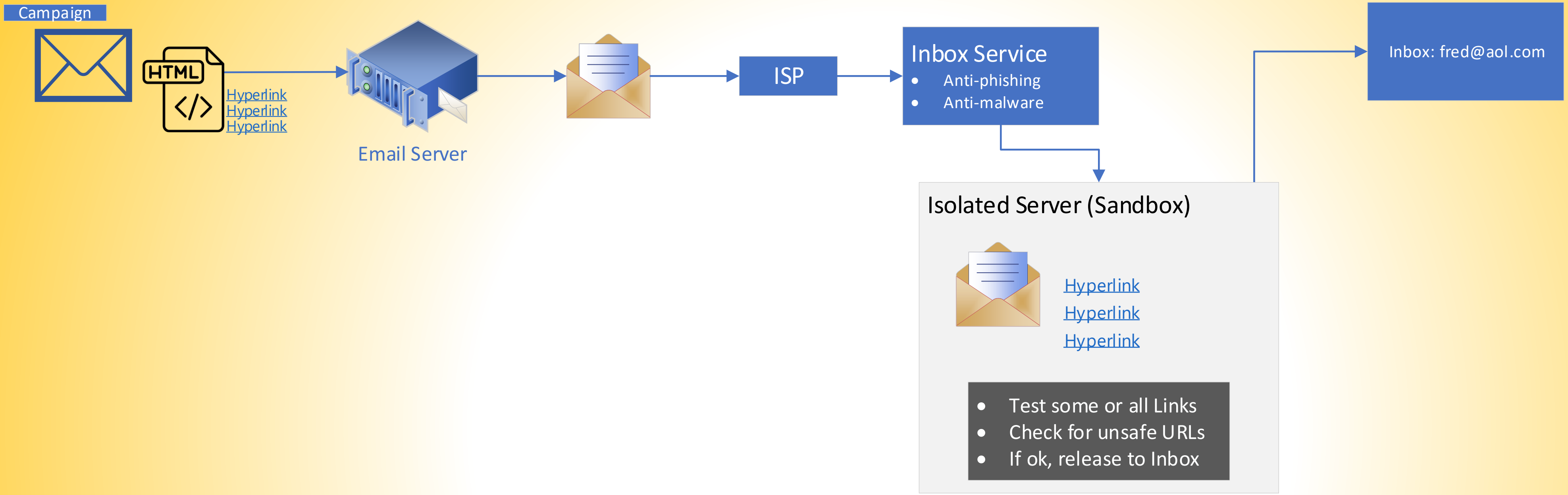
- After the contact is redirected to the proper URL, the click request is saved to the reporting system
- The tracking link contains the ID of the contact so we know who clicked on what links
- The click reporting is a good indicator of how well the campaign performed

Inbox



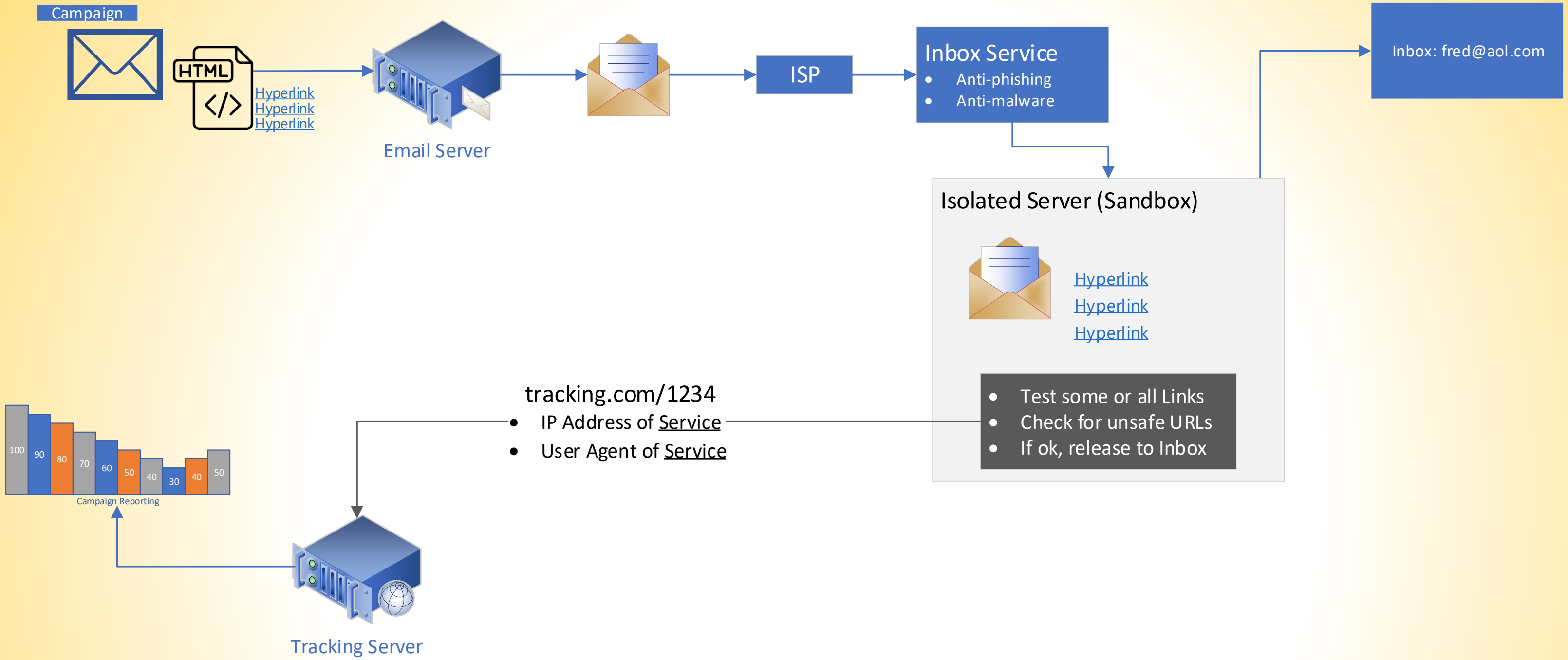
- The contact’s email message is sent to the contact’s ISP (i.e. gmail.com)
- For most contacts, the ISP immediately places the new message in the contact’s inbox
- The contact sees the message without any delays

Inbox Services



- But some ISPs allow companies to add a service that looks at the message before it is placed in the contact's inbox
- These services can be used to check if the email is safe
  - Open the email on an isolated virtual server
  - Click on some or all of the links and check for unsafe patterns and Phishing attacks
  - If ok, release the message to the contact's inbox
- These services will want to evaluate the message as soon as it comes in so the contact doesn't see any noticeable delay
  - The immediate click response is a key feature of BOTs
  - But an actual contact may also have this fast response if the email comes in while they are at the computer

Clicks from an Inbox Service



ISP

Inbox Service
- Anti-phishing
- Anti-malware

Inbox: fred@aol.com

Isolated Server (Sandbox)

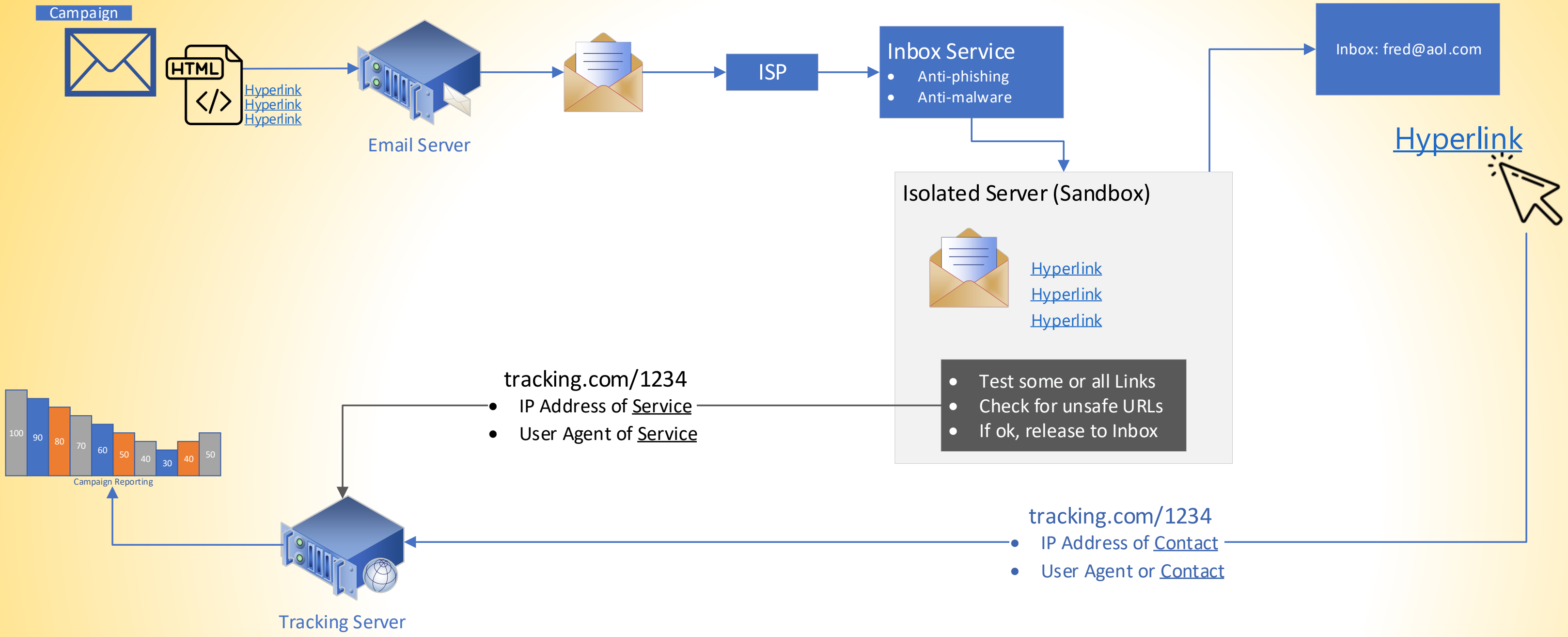
Hyperlink
Hyperlink
Hyperlink

- Test some or all Links
- Check for unsafe URLs
- If ok, release to Inbox

tracking.com/1234
- IP Address of Service
- User Agent of Service

- But the clicks from these inbox services look the same as normal clicks to the tracking server
- In fact, these services want to hide their existence as much as possible
  - If a malicious email can detect the click is coming from a protective service, they could return a safe link to get past the service checks
  - Then the malicious site could change the redirect to an unsafe destination for the actual contact

Organic Click



Email Server

ISP

Inbox Service
- Anti-phishing
- Anti-malware

Inbox: fred@aol.com

Hyperlink

Isolated Server (Sandbox)

Hyperlink
Hyperlink
Hyperlink

- Test some or all Links
- Check for unsafe URLs
- If ok, release to Inbox

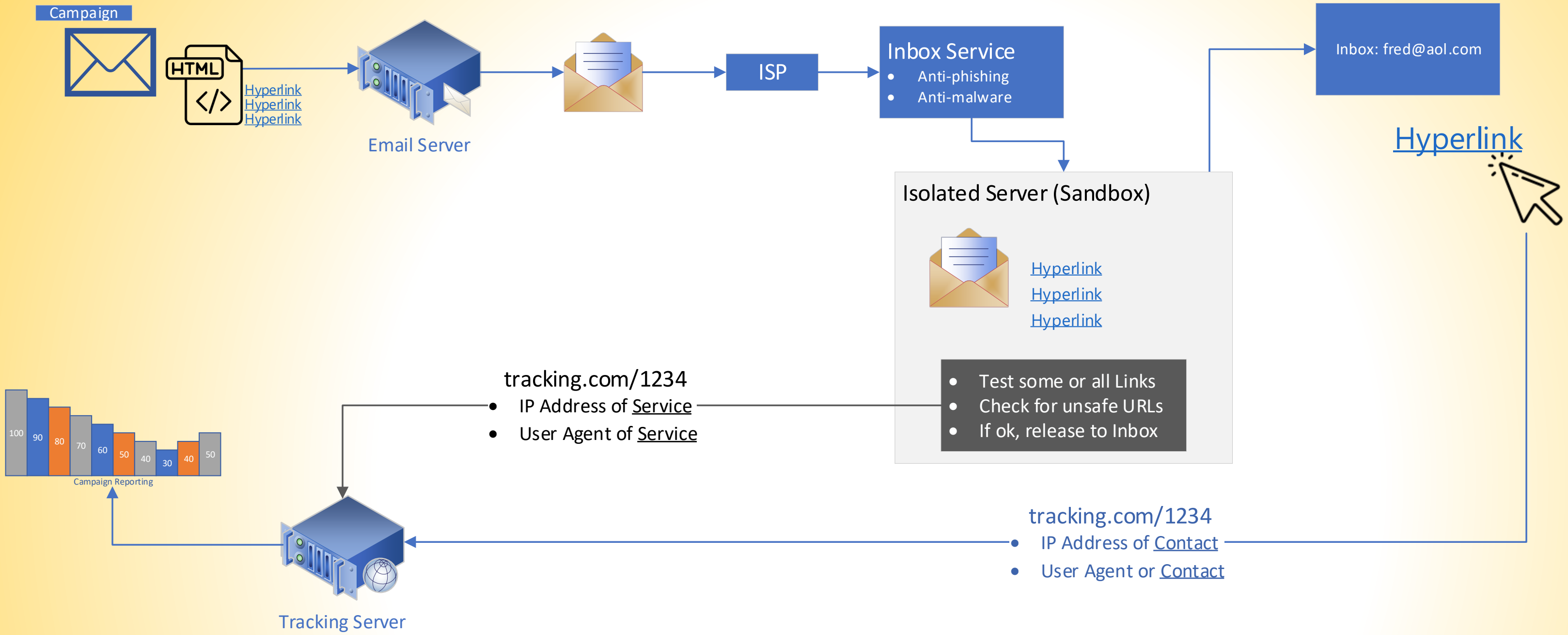
tracking.com/1234
- IP Address of Service
- User Agent of Service

tracking.com/1234
- IP Address of Contact
- User Agent or Contact

Tracking Server

- A contact can still organically click on a link in the message
- The data in the tracking link will be the same for both sources
- The only difference between the 2 click requests is the IP address and User Agent

Reporting



Campaign

HTML

Hyperlink
Hyperlink
Hyperlink

Email Server

ISP

Inbox Service
- Anti-phishing
- Anti-malware

Inbox: fred@aol.com

Hyperlink

Isolated Server (Sandbox)

Hyperlink
Hyperlink
Hyperlink

- Test some or all Links
- Check for unsafe URLs
- If ok, release to Inbox

tracking.com/1234
- IP Address of Service
- User Agent of Service

tracking.com/1234
- IP Address of Contact
- User Agent or Contact

Tracking Server

- Since the inbox service is beneficial, we don’t want to alter the behavior the of the click request
- But if we can detect if the click came from an inbox service, we can discard or flag the click request to allow reporting to be more accurate

IP Address and User Agent

- The IP address is like a phone number for the internet
  - It tells use where the request is coming from and allow use to send back a response to that same computer
- The User Agent string is a way to determine the type of browser and/or device the request is coming from
  - The User Agent string for an iPhone is different from the UA from Outlook on Windows 10
  - This allows web sites to adjust the HTML to fit the device better
  - The UA also changes with installed features like Flash
  - 100 different computers may have 40 to 60 unique User Agent strings
- The combination of these 2 values gives us a good indication of who is making the click request

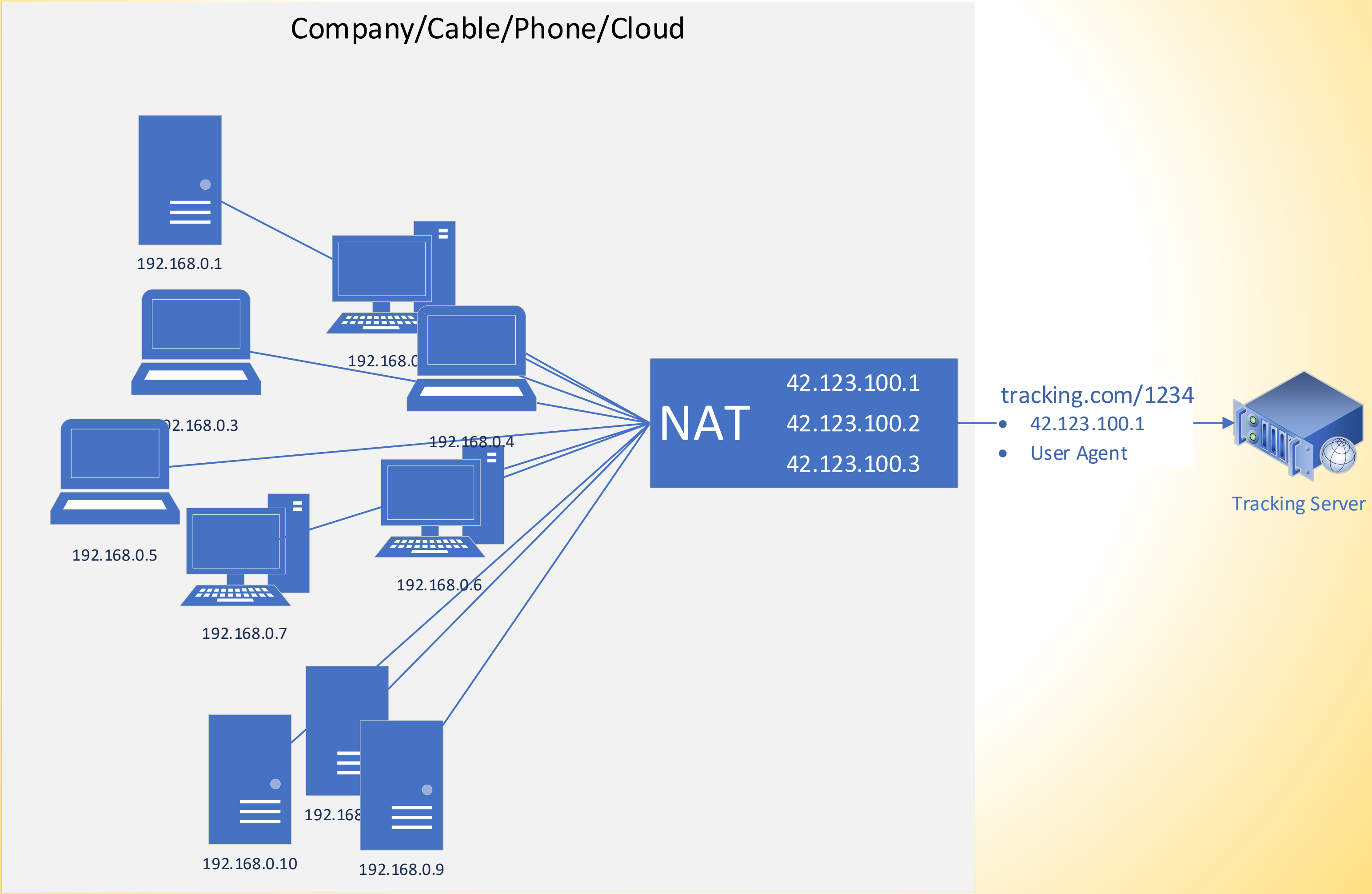
IP Address Issues

- It is not easy to fake an IP address since the response needs to be able to get back to the requesting computer
- But there are still many ways to hide you real IP address
  - Tor networks allows you to submit your request via another host hiding your real IP address
  - Multiple request could each have their own IP address
- But even non-malicious request obfuscate the IP address
- The primary problem comes from Network address Translation (NAT)

Network Address Translation (NAT)

- Natting is use by virtually all companies
- Computers owned by a company do not each have their own pubic IP address
- The company builds their own internal network
  - Each computer gets a private IP address
  - The NAT has 3 or more public IP addresses for all outgoing requests
  - So requests from any computer within the company reuse these public IPs
  - 100 computers may only use 3 or 6 public IP address
- Natting is used beyond just companies
  - Cable providers share public IPs for multiple homes
  - Phone companies share public IPs for requests from smart phones
  - Cloud service providers like Amazon Web Servers (AWS) can share IP address across multiple account

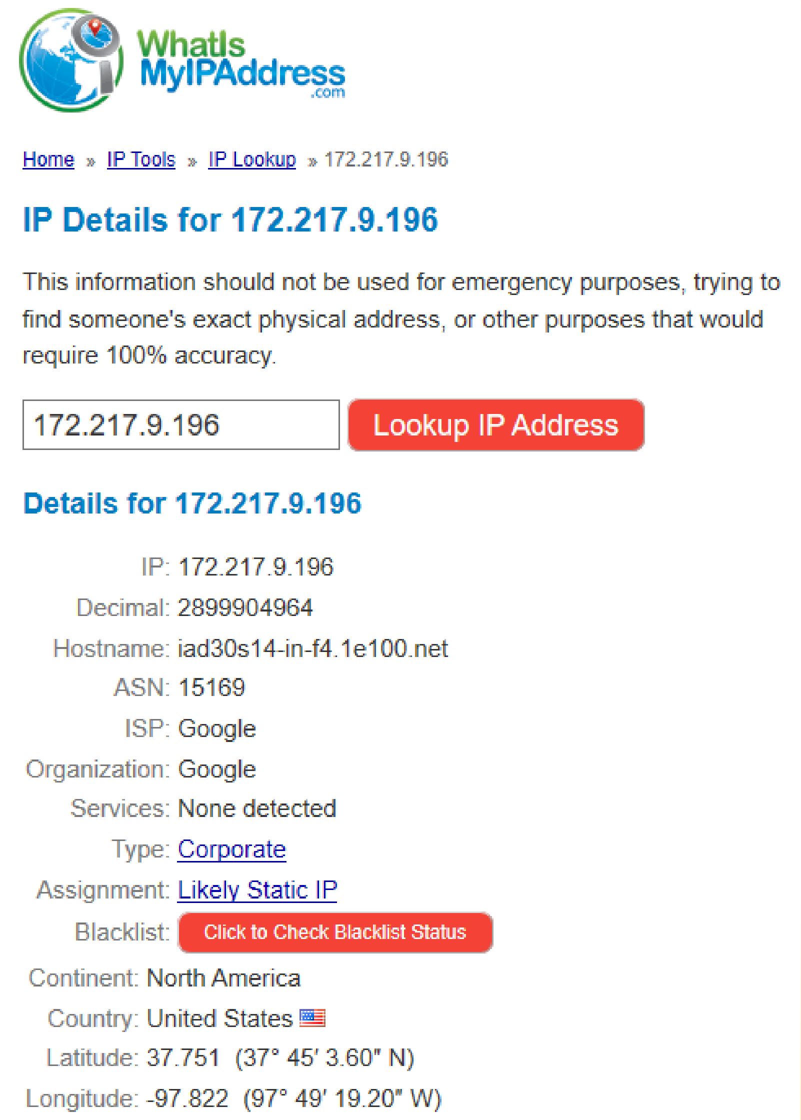
Network Address Translation (NAT)



IP Ownership

- There are a limited number of IP address available (IPv4 = 4.3 billion)
- Most IP address are obtained in blocks
- A block of IP addresses is called a CIDR range
  - Classless Inter-Domain Routing
  - https://en.wikipedia.org/wiki/Classless_Inter-Domain_Routing
- CIDR ranges are associated with an AS Number
  - Autonomous system
  - https://en.wikipedia.org/wiki/Autonomous_system_(Internet)
- AS Numbers are owned by a single company identified by their AS Name

IP Ownership



Home » IP Tools » IP Lookup » 172.217.9.196

IP Details for 172.217.9.196

This information should not be used for emergency purposes, trying to find someone's exact physical address, or other purposes that would require 100% accuracy.

172.217.9.196 Lookup IP Address

Details for 172.217.9.196

IP:	172.217.9.196
Decimal:	2899904964
Hostname:	iad30s14-in-f4.1e100.net
ASN:	15169
ISP:	Google
Organization:	Google
Services:	None detected
Type:	Corporate
Assignment:	Likely Static IP
Blacklist:	Click to Check Blacklist Status
Continent:	North America
Country:	United States
Latitude:	37.751 (37° 45′ 3.60″ N)
Longitude:	-97.822 (97° 49′ 19.20″ W)

- There are services to lookup ownership of an IP address
- These lookups can tell you the owner of the IP (AS Name or ASN)
- They can also tell us lat and long of the location of the IP owner
- There are also databases that can map IP to CIDR ranges, AS Numbers and Names
- This is very helpful for cloud and mobile phone companies
- For example, if you have access to a CIDR database, you can map thousands of IPs to one of the 2 mains datacenters for Amazon Web Services

BOT Detection

- Historically, blacklisting a few known BOT IP addresses and User Agent strings has been adequate
  - Frequently a relatively small set of IP addresses were used, perhaps with just one or two unique UAs
- But the newer protective services are often hosted on cloud services and can have hundreds of IPs available
- They also tend to use User Agents that are very common (iPhone, Chrome on Windows, etc.)
- So a different approach is needed

Sessionization

- Sessionization is an approach for linking request together
- Most email messages contain multiple links
- If an inbox service clicks on 10 of these links one after the other, it would be helpful to group these click requests together into a single session
- The session should be based on requests coming from the same computer
- A session could also be based on the same Inbox (same contact)
- By grouping the click requests from a common source, we extend the number of features we can look at
  - Number of clicks in the session
  - Number of unique links in the session
  - Number of contacts in the session

Sessionization

- Sessionization requires 2 things
  - Ordering of the click requests, i.e. click date/time
  - Grouping Parameters
    - Same IP and User Agent
    - Same IP
    - Same UA
    - Same CIDR
    - Same Contact
    - Same IP, UA and Contact
    - Total duration between first click and last click
    - Max time between click times
- The ordering by click date is straight forward
- But the grouping can have many different approaches

Sessionization

- The approach taken for sessionization was:
  - If time between exceeds 120 second, a new session is created
  - 3 grouping option were are looked. If the first yields good results, we can skip the other 2 grouping options
    - Group by both IP and User Agent
    - Group by IP only
    - Group by InboxID (email contact)
- Since historically, the combination of IP and UA has been successful, it was the first grouping strategy tried

Aggregation

- Since we are taking a session based approach, most of our metrics used in our modeling will be aggregate data
- Some of the aggregation options used include:
  - Number of InboxIDs (# of contacts)
  - Total number of click requests
  - Total number of email domains (i.e. gmail.com, yahoo.com, etc.)
  - Unique Links requested
    - 100 InboxIDs with 7 clicks each on a message with 40 links, are they all the same 7 links?
  - Mean duration between SendDate and ClickDate

Approach

- Since there are no labels with the click requests, unsupervised learning is the only option
- Validation would need to be done manually
  - Randomly pick n sessions for each group discovered
  - Do research (IP Lookups, MX Lookups, AWS) to determine if click was from a BOT
  - Manually add labels to the unsupervised groups
- Apply labeled sessions to raw data
  - Join sessionized data to raw data on grouping values (i.e. IP and UA)
  - Append labels to the raw data
- Derive a score
  - Some sessions from the same IP/UA may not have been labeled the same
  - % of raw requests labeled as BOT requests
- Use score to blacklist IP/UA and/or CIDR ranges

Possible Approaches to use in Production

- Sample click requests on a weekly/monthly basis
    - Identity BOTs IP/UA or CIDR
    - Add to existing blacklist
- Extract parameters used to differentiate the different unsupervised labels
    - Add caching layer to tracking server
    - Load session data into cache
    - Use model parameters to determine BOT label