



Zurich Research Laboratory
Tokyo Research Laboratory



Relationship Discovery with NetFlow to Enable Business-Driven IT Management

Andreas Kind <ank@zurich.ibm.com>
Dieter Gantenbein <dga@zurich.ibm.com>
Hiroaki Etoh <etoh@jp.ibm.com>

IBM | Apr 06 | Relationship Discovery

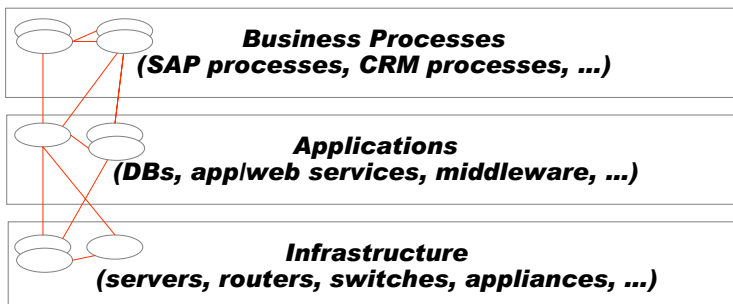
www.zurich.ibm.com/aurora

Zurich Research Laboratory



Business-Driven IT Management

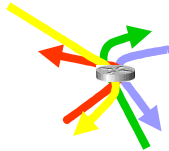
- Manage IT according to high-level business goals
- Discover relationships between all IT layers



- This talk is about using NetFlow for automatic relationship discovery

What is NetFlow?

- The “phone bill” for traffic in IP networks
- How much is who talking to whom over what protocols with which applications?

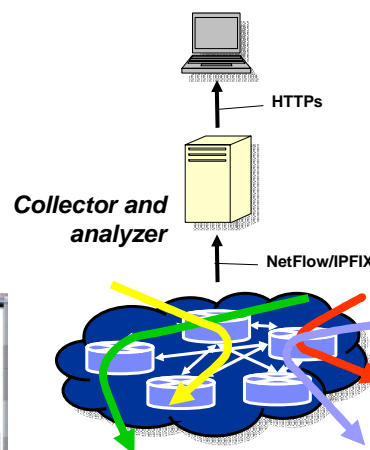


9.4.68.175	22	9.4.71.177	3476	6/tcp	1301	381278	...
9.4.71.177	3476	9.4.68.175	22	6/tcp	1298	128750	...
9.4.64.246	0	224.0.0.13	0	103/pim	45	2880	...
9.4.68.196	33496	9.4.71.255	111	17/udp	15	2310	...
...							...

- Flow definition
 - A flow is a set of 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 derived from the data contained in the packet and from the packet treatment at the observation point
- Flow-based network profiling differs from monitoring of network and system devices
 - State information is polled with device monitoring typically
 - Application flow information is pushed with profiling

What is NetFlow?

- Primary technology for ...
 - Network accounting
 - Bandwidth usage analysis
 - Network anomaly-detection
- Emerging technology for traffic engineering and capacity planning



Flows and Flow Events

Flow

$f = (src, dst, proto, srv, rcvd_flag)$ in F

Flow Event

$e = (f, t_s, t_e, octs, pkts)$ in E

Flow Events

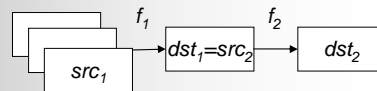
$E(f) = \{ e \text{ in } E \mid f_e = f \}$

Flow predicate

$f_p: F \times F \rightarrow \{1, 0\}$

For example:

$$f_p = \begin{cases} 1 & \text{if } (dst_1 = src_2) \text{ and } (src_1 \neq dst_2) \\ 0 & \text{otherwise} \end{cases}$$



With: $f_1 = (src_1, dst_1, proto_1, srv_1, rcvd_flag_1)$
 $f_2 = (src_2, dst_2, proto_2, srv_2, rcvd_flag_2)$

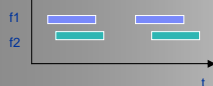
Flow Event Pairs

Flow event pairs

$P(f_1, f_2) = \{ (e_1, e_2) \mid \begin{array}{l} 0 \leq t_s(e_2) - t_s(e_1) < t_{max} \\ f_p(f(e_1), f(e_2)) = 1 \end{array} \text{ and} \right.$
 with
 $\left. e_1 \text{ in } E(f_1) \text{ and } e_2 \text{ in } E(f_2) \right\}$

$t_{max} = 10s$

Example:



$|E| = 4$
 $|P(f_1, f_2)| = 2$

Flow Correlation

Flow event correlation value

$$c(f_1, f_2) = \sum D(t_s(e_2) - t_s(e_1)) / |P(f_1, f_2)|$$

with (e_1, e_2) in $P(f_1, f_2)$

Distribution

$D(0) = 1.00$
 $D(1) = 0.99$
 $D(2) = 0.98$
 ...
 $D(9) = 0.01$



Examples

$c = 1.97/2 = \mathbf{0.985}$ $|E| = 3, |\Sigma P| = 3$ $c = 0.99/1 = \mathbf{0.99}$



Flow Correlation Confidence

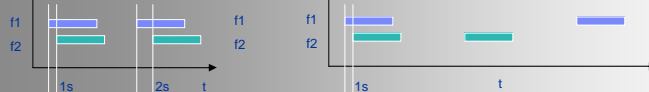
Confidence value

$$c'(f_1, f_2) = \underbrace{(1 - 1/(|P(f_1, f_2)| + 1))}_{(1)} * \underbrace{|P(f_1, f_2)|/|E|}_{(2)}$$

- (1): The more pairs the better
But what if we have MANY events?
- (2): The higher the #pairs relative to #events the better

Examples

$c = 1.97/2 = \mathbf{0.985}$ $c = 0.99/1 = \mathbf{0.99}$ *Similar correlation value*
 $c' = 0.66 * 1 = \mathbf{0.66}$ $c' = 0.5 * 0.5 = \mathbf{0.25}$ *But lower confidence value*



Implementation

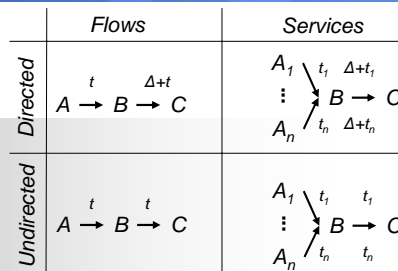
1. Parse NetFlow records and sort flow events by start times
2. Walk through sorted flow events with window of t_{max} and update correlation values of flow pairs which occur as flow event pairs in time window
3. Compute confidence values for all flow pairs identified
4. Sort identified flow pairs by $c * c'$
5. Generate output with top flow correlations in XML
6. Merge last 24 hourly correlation files into a daily correlation file; merge last 30 daily correlation files into a monthly correlation file

XML Output

```
<?xml version="1.0" ?>
<flow-correlation>
...
<flows ids="753" total="5301">
  <counters>
    <packets>314730</packets>
    <octets>9323000</octets>
  </counters>
  <flow id="581" proto="6" appl="23"
    service="873" tos="0x00">
    <endpoint type="src" address="*"
      port="873" domain="K Bld" />
    <endpoint type="dst" address="9.4.4.138"
      port="0" domain="AIX Srv" />
    <counters>
      <flows>1</flows>
      <packets>7</packets>
      <octets>542</octets>
    </counters>
  </flow>
  <flow id="2" proto="6" appl="16"
    service="730" tos="0x00">
    <endpoint type="src" address="9.4.20.44"
      port="730" domain="K Bld" />
    <endpoint type="dst" address="9.4.9.135"
      port="0" domain="AIX Srv" />
    <counters>
      <flows>1</flows>
      <packets>7</packets>
      <octets>542</octets>
    </counters>
  </flow>
</flows>
<correlations type="undirected" total="14">
  <correlation srcid="9" dstid="2">
    value="71" confidence="36"/>
  <correlation srcid="182" dstid="2">
    value="33" confidence="54"/>
  ...
</correlations>
<correlations type="directed" total="22">
  <correlation srcid="25" dstid="176">
    value="88" confidence="34"/>
  <correlation srcid="11" dstid="47">
    value="27" confidence="14"/>
  ...
</correlations>
</flow-correlation>
```

Flows and services

Flow correlations (directed/undirected)

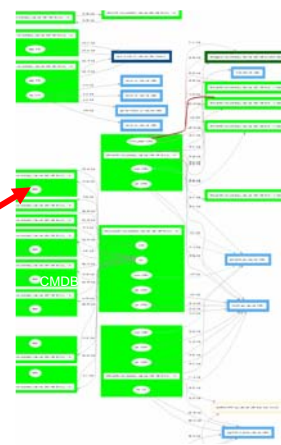
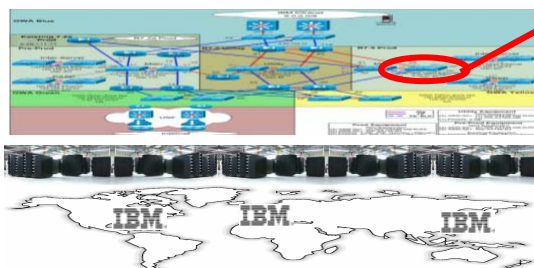


How Do Relationships Enable BDIM?

- Example 1:
 - Problem statement:** Identify opportunities for application co-location on single servers.
 - Business objective:** More accurate transition plans reduce actual transition costs. Smaller number of servers. Reduced future operation costs.
 - Action based on relationship information:** Suggest co-location of applications onto single server according to service dependencies and acceptable server loads.
- Example 2:
 - Problem statement:** Find optimal staging of server relocations.
 - Business objective:** Cut relocation costs.
 - Action based on relationship information:** Schedule relocations for minimal business impact.

Application in CRM Siebel Production Environment

- Relationship discovery revealed traffic, middleware, and application relationships
- Discovered relationships:
 - Siebel server dependency from backup service
 - Application server dependency from database
 - Web server dependency from authentication directory
 - CRM business process dependency from IT assets X1, X2, ...



Thanks!

<http://www.zurich.ibm.com/aurora>