Cloud-Scale BGP and NetFlow Analysis

Jim Frey, VP Product, Kentik Technologies
December 15, 2015
Agenda

• Common NetOps Stress points
• Helpful Data Sets - NetFlow, BGP
• Handling NetFlow and BGP at Cloud Scale
• Kentik’s Approach
• Wrap-Up / Q&A
NetOps Stress Points: Needing *Instant* Answers

**Things You Need Answers to About/From Your Network**

- Where in my network is the problem?
- Is this an attack or legitimate traffic?
- Does performance meet expectations?
- How should I allocate my resources in the future?

---

**NetOps Stress Points Diagram**

- **X**
- **S**
- **R**
- **$$$**

---

© kentik
What We Hear....

To Address These Questions, NetOps Needs:

- Accurate Visibility, Without Delay
- Relevant Alerts: No False Positives or Negatives
- Complete Data: Breadth + Depth
- Fast/Flexible Data Exploration
- Tools that don’t suck (time or $$)
What Data Sets Can Help?

*And which ones can do the job cost effectively?*
Primary Network Monitoring Data Choices

### Polled Stats

**Examples**
- SNMP, WMI

**Advantages**
- Ubiquitous
- Good for monitoring device health/status/activity

**Disadvantages**
- No traffic detail
- Typically no frequent than every 5 minutes truly anti-real-time

### Flow Records

**Examples**
- NetFlow, sFlow, IPFIX

**Advantages**
- Details on traffic src/dest/content, etc.
- Very cost effective

**Disadvantages**
- NRT (near real-time) at best
- Incomplete app-layer detail
- Limited performance metrics
- Data volumes can be massive

### Packet Inspection

**Examples**
- Packets -> xFlow
- Long term stream-to-disk

**Advantages**
- Most complete app layer detail
- True real-time (millisecond lvl)
- Complete vendor independent

**Disadvantages**
- Expensive to deploy at scale
- Requires network tap or SPAN
- Packet captures can be massive
### Secondary Network Monitoring Data Choices

#### Log Records

**Examples**
- Syslog

**Advantages**
- Continuous/streaming
- Unique, device-specific info
- True real-time

**Disadvantages**
- No standards – must have very flexible search/mapping tools
- Data volumes can be massive

#### Routing/Path Data

**Examples**
- OSPF, IGRP, BGP

**Advantages**
- Details on traffic paths and provider volumes
- Insights into Internet factors

**Disadvantages**
- Address data only – no awareness of traffic
- Must peer with routers to get updates

#### Synthetic Agents

**Examples**
- IP SLA, Independent test sw

**Advantages**
- Assess functions/services 24x7
- Provides both availability and performance measures

**Disadvantages**
- Deploying/maintaining enough agents to achieve full coverage
- Only an approximation of real user experience (at best)
Key Assertion:

Use Multiple Data Types for Best Results

• You never know which data set will present the specific insights you need
• The challenge (real magic) comes from correlating multiple datasets, i.e.:
  • Behavioral observations with configuration changes
  • Trends with underlying traffic details
  • Routing data with traffic data
Why Correlate Routing Data with Traffic Data?

For Providers
- Recognizing new service opportunities based on subscriber (and peer) behavior
- Optimizing peering relationships for cost control

For Web Services / Commerce
- Recognizing where your customers are and how they reach you
- Managing peering relationships for best customer experience

For Enterprise
- Assessing how your connectivity providers perform/compare
- Building Internet IQ – how you connect/relate to the outside world
Cloud Scale for NetFlow and BGP: The Big Data Challenge

Why can’t we just use our existing tools?
Network traffic has grown exponentially; Legacy tools/tech haven’t kept pace.

Result? Fragmented tools, visibility gaps, 
unanswered questions.
Why Big Data?

- Network Monitoring Data is Big Data
  - Meets Volume/Variety/Velocity Test
  - Billions of records/day (millions/second)
- Big Data architectures are considered best practices today for open/flexible correlation, analytics
Why Big Data?

- Network Monitoring Data *IS* Big Data
  - Meets Volume/ Variety/ Velocity Test
  - Billions of records/day (millions/ second)
- Big Data architectures are considered best practices today for open/ flexible correlation, analytics

Specific Challenges For NetFlow + BGP

**Existing solutions shortfalls:**
- Flexibility for moving between viewpoints and into full details
- Data Completeness due to reliance on summarized/ aggregated flow data
- Speed: Generating new analysis in a timely manner
How to Get/Use Big Data Approach?
How to Get/Use Big Data Approach?

1. BYO - Build Your Own

• Pick back end & reporting/analysis tools (open source = free?)
• Procure operating platforms (hard, virtual, or cloud servers = $$)
• Integrate, add data sources, and get it up and running (dev = $$)
• Keep it up and running (ops/admin = $$)
How to Get/Use Big Data Approach?

1. BYO - Build Your Own
   • Pick back end & reporting/analysis tools (open source = free?)
   • Procure operating platforms (hard, virtual, or cloud servers = $$)
   • Integrate, add data sources, and get it up and running (dev = $$)
   • Keep it up and running (ops/admin = $$)

2. Let SOMEONE ELSE build/optimize/operate
   • Subscribe to SaaS (ops $$)
   • Just Send Your Data and enjoy the ride!
Kentik’s Answer

How we address the Big Data challenge to meet the needs of Network Operators now
Kentik Detect: the first and only SaaS Solution
For Network Ops Management & Visibility at Terabit Scale

The Network is the Sensor

Big Data Network Telemetry Platform

Kentik Data Engine

Web Portal
Real-time & historical queries

Alerts
E-mail / Syslog / JSON

Open API
SQL / RESTful

Analyze & Take Action

Cloud-Based
Real-Time
Multi-Tenant
Open
Global
What’s Behind the Kentik Data Engine

Multi-tiered/Clustered for Scale / Load Balancing / HA, Hosted by Kentik

Optimized for Massive Data Ingest & Rapid Query Response
Kentik Portal Dashboard
Top Traffic Flows
Traffic by Source Geography

<table>
<thead>
<tr>
<th>src_geo</th>
<th>Avg Mb/sec</th>
<th>Percent Total</th>
<th>99th Percentile</th>
<th>Max Mb/sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>NL</td>
<td>41.17</td>
<td>85</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>DE</td>
<td>11.01</td>
<td>23</td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>UA</td>
<td>9.79</td>
<td>93</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td>8.22</td>
<td>18</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>JP</td>
<td>5.34</td>
<td>11</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>KR</td>
<td>3.48</td>
<td>10</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>FR</td>
<td>2.75</td>
<td>19</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>PL</td>
<td>2.39</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>ES</td>
<td>1.59</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>PT</td>
<td>1.31</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>GB</td>
<td>1.48</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>AU</td>
<td>1.02</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>BR</td>
<td>0.99</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>NO</td>
<td>0.93</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>
AS Path Changes

[Graph and data table from the Kentik dashboard showing bandwidth usage over time for various AS paths.]
AS Top Talkers and Drill Down Options
Peering Analytics: ASN by Dest Country Paths
Peering Analytics: Traffic by BGP Paths
Peering Analytics: Traffic by Origin AS (“Last Hop”)

Graph showing traffic flow by Origin AS and destination country.
Peering Analytics: Traffic by Transit AS
Key Takeaways: Cloud Scale NetFlow + BGP

Why You Need It
- Clear Insight into external/Internet network traffic behaviors
- Improved customer/subscriber engagement
- Reduced network operating costs

Technical Path to Success
- This is a big data problem, requiring high capacity/speed for data management, correlation, exploration, and analytics
- SaaS solutions are a fully viable option
Thank You!

Jim Frey
VP Product
Kentik Technologies
jfrey@kentik.com
@jfrey80