BGP Route Aggregation
Best Practices

Philip Smith

APNIC 23
28th February
Bali, Indonesia
Agenda

- What is Aggregation?
- RIPE-399 Aggregation Recommendations
- What is happening world wide?
Aggregation

- Aggregation means announcing the address block received from the RIR to the other ASes connected to your network.
- Subprefixes of address block must NOT be announced to Internet unless aiding traffic engineering for multihoming.
- Subprefixes of this aggregate will be present internally in the ISP network.
Announcing an Aggregate

- ISPs who don’t and won’t aggregate are held in poor regard by community
- Registries publish their minimum allocation size
  - Anything from a /20 to a /22 depending on RIR
  - Different sizes for different address blocks
- No real reason to see anything longer than a /22 prefix in the Internet
  - BUT there are currently >110000 /24s!
Customer has /23 network assigned from AS100’s /19 address block

AS100 announces customers’ individual networks to the Internet
Aggregation – Bad Example

- Customer link goes down
  - Their /23 network becomes unreachable
  - /23 is withdrawn from AS100’s iBGP
- Their ISP doesn’t aggregate its /19 network block
  - /23 network withdrawal announced to peers
  - starts rippling through the Internet
  - added load on all Internet backbone routers as network is removed from routing table

- Customer link returns
  - Their /23 network is now visible to their ISP
  - Their /23 network is re-advertised to peers
  - Starts rippling through Internet
  - Load on Internet backbone routers as network is reinserted into routing table
  - Some ISP’s suppress the flaps
  - Internet may take 10-20 min or longer to be visible
  - Where is the Quality of Service???
Customer has /23 network assigned from AS100’s /19 address block

AS100 announced /19 aggregate to the Internet
Aggregation – Good Example

- Customer link goes down
  - their /23 network becomes unreachable
  - /23 is withdrawn from AS100’s iBGP
- /19 aggregate is still being announced
  - no BGP hold down problems
  - no BGP propagation delays
  - no damping by other ISPs

- Customer link returns
- Their /23 network is visible again
  - The /23 is re-injected into AS100’s iBGP
- The whole Internet becomes visible immediately
- Customer has Quality of Service perception
Aggregation – Summary

- Good example is what everyone should do!
  - Adds to Internet stability
  - Reduces size of routing table
  - Reduces routing churn
  - Improves Internet QoS for everyone

- Bad example is what too many still do!
  - Why? Lack of knowledge?
  - Laziness?
The Internet Today
(January 2007)

- Current Internet Routing Table Statistics
  - BGP Routing Table Entries: 207115
  - Prefixes after maximum aggregation: 112059
  - Unique prefixes in Internet: 100861
  - Prefixes smaller than registry alloc: 105377
  - /24s announced: 110473
  - only 5748 /24s are from 192.0.0.0/8
  - ASes in use: 24066
“The New Swamp”

- ‘Swamp Space’ is name used for areas of poor aggregation
  - The original swamp was 192.0.0.0/8 from the former class C block
    - Name given just after the deployment of CIDR
  - The new swamp is creeping across all parts of the Internet
    - Not just RIR space, but “legacy” space too
"The New Swamp"
RIR Space – February 1999

RIR blocks contribute 49393 prefixes or 88% of the Internet Routing Table

<table>
<thead>
<tr>
<th>Block</th>
<th>Networks</th>
<th>Block</th>
<th>Networks</th>
<th>Block</th>
<th>Networks</th>
<th>Block</th>
<th>Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>24/8</td>
<td>165</td>
<td>74/8</td>
<td>0</td>
<td>124/8</td>
<td>0</td>
<td>205/8</td>
<td>2584</td>
</tr>
<tr>
<td>41/8</td>
<td>0</td>
<td>75/8</td>
<td>0</td>
<td>125/8</td>
<td>0</td>
<td>206/8</td>
<td>3127</td>
</tr>
<tr>
<td>58/8</td>
<td>0</td>
<td>76/8</td>
<td>0</td>
<td>126/8</td>
<td>0</td>
<td>207/8</td>
<td>2723</td>
</tr>
<tr>
<td>59/8</td>
<td>0</td>
<td>80/8</td>
<td>0</td>
<td>188/8</td>
<td>0</td>
<td>208/8</td>
<td>2817</td>
</tr>
<tr>
<td>60/8</td>
<td>0</td>
<td>81/8</td>
<td>0</td>
<td>189/8</td>
<td>0</td>
<td>209/8</td>
<td>2574</td>
</tr>
<tr>
<td>61/8</td>
<td>3</td>
<td>82/8</td>
<td>0</td>
<td>190/8</td>
<td>0</td>
<td>210/8</td>
<td>617</td>
</tr>
<tr>
<td>62/8</td>
<td>87</td>
<td>83/8</td>
<td>0</td>
<td>192/8</td>
<td>6275</td>
<td>211/8</td>
<td>0</td>
</tr>
<tr>
<td>63/8</td>
<td>20</td>
<td>84/8</td>
<td>0</td>
<td>193/8</td>
<td>2390</td>
<td>212/8</td>
<td>717</td>
</tr>
<tr>
<td>64/8</td>
<td>0</td>
<td>85/8</td>
<td>0</td>
<td>194/8</td>
<td>2932</td>
<td>213/8</td>
<td>1</td>
</tr>
<tr>
<td>65/8</td>
<td>0</td>
<td>86/8</td>
<td>0</td>
<td>195/8</td>
<td>1338</td>
<td>216/8</td>
<td>943</td>
</tr>
<tr>
<td>66/8</td>
<td>0</td>
<td>87/8</td>
<td>0</td>
<td>196/8</td>
<td>513</td>
<td>217/8</td>
<td>0</td>
</tr>
<tr>
<td>67/8</td>
<td>0</td>
<td>88/8</td>
<td>0</td>
<td>198/8</td>
<td>4034</td>
<td>218/8</td>
<td>0</td>
</tr>
<tr>
<td>68/8</td>
<td>0</td>
<td>89/8</td>
<td>0</td>
<td>199/8</td>
<td>3495</td>
<td>219/8</td>
<td>0</td>
</tr>
<tr>
<td>69/8</td>
<td>0</td>
<td>90/8</td>
<td>0</td>
<td>200/8</td>
<td>1348</td>
<td>220/8</td>
<td>0</td>
</tr>
<tr>
<td>70/8</td>
<td>0</td>
<td>91/8</td>
<td>0</td>
<td>201/8</td>
<td>0</td>
<td>221/8</td>
<td>0</td>
</tr>
<tr>
<td>71/8</td>
<td>0</td>
<td>121/8</td>
<td>0</td>
<td>202/8</td>
<td>2276</td>
<td>222/8</td>
<td>0</td>
</tr>
<tr>
<td>72/8</td>
<td>0</td>
<td>122/8</td>
<td>0</td>
<td>203/8</td>
<td>3622</td>
<td></td>
<td></td>
</tr>
<tr>
<td>73/8</td>
<td>0</td>
<td>123/8</td>
<td>0</td>
<td>204/8</td>
<td>3792</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**“The New Swamp”**

RIR Space – February 2006

RIR blocks contribute 161287 prefixes or 88% of the Internet Routing Table

<table>
<thead>
<tr>
<th>Block</th>
<th>Networks</th>
<th>Block</th>
<th>Networks</th>
<th>Block</th>
<th>Networks</th>
<th>Block</th>
<th>Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>24/8</td>
<td>3001</td>
<td>74/8</td>
<td>109</td>
<td>124/8</td>
<td>292</td>
<td>205/8</td>
<td>2934</td>
</tr>
<tr>
<td>41/8</td>
<td>41</td>
<td>75/8</td>
<td>2</td>
<td>125/8</td>
<td>682</td>
<td>206/8</td>
<td>3879</td>
</tr>
<tr>
<td>58/8</td>
<td>606</td>
<td>76/8</td>
<td>4</td>
<td>126/8</td>
<td>27</td>
<td>207/8</td>
<td>4385</td>
</tr>
<tr>
<td>59/8</td>
<td>628</td>
<td>80/8</td>
<td>1925</td>
<td>188/8</td>
<td>1</td>
<td>208/8</td>
<td>3239</td>
</tr>
<tr>
<td>60/8</td>
<td>468</td>
<td>81/8</td>
<td>1350</td>
<td>189/8</td>
<td>0</td>
<td>209/8</td>
<td>5611</td>
</tr>
<tr>
<td>61/8</td>
<td>2396</td>
<td>82/8</td>
<td>1158</td>
<td>190/8</td>
<td>39</td>
<td>210/8</td>
<td>3908</td>
</tr>
<tr>
<td>62/8</td>
<td>1860</td>
<td>83/8</td>
<td>1130</td>
<td>192/8</td>
<td>6927</td>
<td>211/8</td>
<td>2291</td>
</tr>
<tr>
<td>63/8</td>
<td>2837</td>
<td>84/8</td>
<td>971</td>
<td>193/8</td>
<td>5203</td>
<td>212/8</td>
<td>2920</td>
</tr>
<tr>
<td>64/8</td>
<td>5374</td>
<td>85/8</td>
<td>1426</td>
<td>194/8</td>
<td>4061</td>
<td>213/8</td>
<td>3071</td>
</tr>
<tr>
<td>65/8</td>
<td>3785</td>
<td>86/8</td>
<td>650</td>
<td>195/8</td>
<td>3519</td>
<td>216/8</td>
<td>6893</td>
</tr>
<tr>
<td>66/8</td>
<td>6292</td>
<td>87/8</td>
<td>629</td>
<td>196/8</td>
<td>1264</td>
<td>217/8</td>
<td>2590</td>
</tr>
<tr>
<td>67/8</td>
<td>1832</td>
<td>88/8</td>
<td>328</td>
<td>198/8</td>
<td>4908</td>
<td>218/8</td>
<td>1220</td>
</tr>
<tr>
<td>68/8</td>
<td>3069</td>
<td>89/8</td>
<td>113</td>
<td>199/8</td>
<td>4156</td>
<td>219/8</td>
<td>1003</td>
</tr>
<tr>
<td>69/8</td>
<td>3315</td>
<td>90/8</td>
<td>2</td>
<td>200/8</td>
<td>6757</td>
<td>220/8</td>
<td>1657</td>
</tr>
<tr>
<td>70/8</td>
<td>1597</td>
<td>91/8</td>
<td>2</td>
<td>201/8</td>
<td>1614</td>
<td>221/8</td>
<td>765</td>
</tr>
<tr>
<td>71/8</td>
<td>888</td>
<td>121/8</td>
<td>0</td>
<td>202/8</td>
<td>9759</td>
<td>222/8</td>
<td>914</td>
</tr>
<tr>
<td>72/8</td>
<td>1772</td>
<td>122/8</td>
<td>0</td>
<td>203/8</td>
<td>9527</td>
<td></td>
<td></td>
</tr>
<tr>
<td>73/8</td>
<td>274</td>
<td>123/8</td>
<td>0</td>
<td>204/8</td>
<td>5474</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
"The New Swamp"

Summary

- RIR space shows creeping deaggregation
  - Today an RIR /8 block averages around 6000 prefixes once fully allocated
  - Existing 74 /8s will eventually cause 444000 prefix announcements

- Food for thought:
  - Remaining 58 unallocated /8s and the 74 RIR /8s combined will cause:
    - 852000 prefixes with 6000 prefixes per /8 density
    - Plus 12% due to “non RIR space deaggregation”
    - → Routing Table size of 954240 prefixes
“The New Swamp” Summary

- Rest of address space is showing similar deaggregation too 😞
- What are the reasons?
  - Main justification is traffic engineering
- Real reasons are:
  - Lack of knowledge
  - Laziness
  - Deliberate & knowing actions
BGP Report
(bgp.potaroo.net)

- 199336 total announcements in October 2006
- 129795 prefixes
  - After aggregating including full AS PATH info
    - i.e. including each ASN’s traffic engineering
  - 35% saving possible
- 109034 prefixes
  - After aggregating by Origin AS
    - i.e. ignoring each ASN’s traffic engineering
  - 10% saving possible
The excuses

- Traffic engineering causes 10% of the Internet Routing table
- Deliberate deaggregation causes 35% of the Internet Routing table
Efforts to improve aggregation

- The CIDR Report
  - Initiated and operated for many years by Tony Bates
  - Now combined with Geoff Huston’s routing analysis
    - www.cidr-report.org
  - Results e-mailed on a weekly basis to most operations lists around the world
  - Lists the top 30 service providers who could do better at aggregating
The CIDR Report

- Also computes the size of the routing table assuming ISPs performed optimal aggregation
- Website allows searches and computations of aggregation to be made on a per AS basis
  - Flexible and powerful tool to aid ISPs
  - Intended to show how greater efficiency in terms of BGP table size can be obtained without loss of routing and policy information
  - Shows what forms of origin AS aggregation could be performed and the potential benefit of such actions to the total table size
  - Very effectively challenges the traffic engineering excuse
Agenda

- What is Aggregation?
- RIPE-399 Aggregation Recommendations
- What is happening world wide?
Route Aggregation Recommendations

- LINX started with aggregation policy for members
  - It failed — “IXP interfering with members business practices”
  - Even though most members voted for policy!
- RIPE Routing Working Group work item from early 2006
  - Based on early LINX concept
  - Authored by Philip Smith, Mike Hughes (LINX CTO) and Rob Evans (UKERNA)
Route Aggregation Recommendations

- RIPE Document — RIPE-399
  - http://www.ripe.net/ripe/docs/ripe-399.html

- Discusses:
  - History of aggregation
  - Causes of de-aggregation
  - Impacts on global routing system
  - Available Solutions
  - Recommendations for ISPs
History:

- Classful to classless migration
  - Clean-up efforts in 192/8
- CIDR Report
  - Started by Tony Bates to encourage adoption of CIDR & aggregation
  - Mostly ignored through late 90s
  - Now part of extensive BGP table analysis by Geoff Huston
- Introduction of Regional Internet Registry system and PA address space
Deaggregation:
Claimed causes (1):

- Routing System Security
  - “Announcing /24s means that no one else can DOS the network”

- Reduction of DOS attacks & miscreant activities
  - “Announcing only address space in use as rest attracts ‘noise’”

- Commercial Reasons
  - “Mind your own business”
Deaggregation:
Claimed causes (2):

- Leakage of iBGP outside of local AS
  - eBGP is NOT iBGP - how many ISPs know this?
- Traffic Engineering for Multihoming
  - Spraying out /24s hoping it will work
  - Rather than being sparing
- Legacy Assignments
  - “All those pre-RIR assignments are to blame”
  - In reality it is both RIR and legacy assignments
Impacts (1):

- **Router memory**
  - Shortens router life time as vendors underestimate memory growth requirements
  - Depreciation life-cycle shortened
  - Increased costs for ISP and customers

- **Router processing power**
  - Processors are underpowered as vendors underestimate CPU requirement
  - Depreciation life-cycle shortened
  - Increased costs for ISP and customers
Impacts (2):

- **Routing System convergence**
  - Larger routing table → slowed convergence
  - Can be improved by faster control plane processors — see earlier

- **Network Performance & Stability**
  - Slowed convergence → slowed recovery from failure
  - Slowed recovery → longer downtime
  - Longer downtime → unhappy customers
Solutions (1):

- CIDR Report
  - Global aggregation efforts
  - Running since 1994
- Routing Table Report
  - Per RIR region aggregation efforts
  - Running since 1999
- Filtering recommendations
  - Training, tutorials, Project Cymru,...
- “CIDR Police”
Solutions (2):

- BGP Features:
  - NO_EXPORT Community
  - NOPEER Community
    - RFC3765 — but no one has implemented it
  - AS_PATHLIMIT attribute
    - Still working through IETF IDR Working Group
  - Provider Specific Communities
    - Some ISPs use them; most do not
Recommendations:

- Announcement of initial allocation as a single entity
- Subsequent allocations aggregated if they are contiguous and bit-wise aligned
- Prudent subdivision of aggregates for Multihoming
- Use BGP enhancements already discussed
- (Oh, and all this applies to IPv6 too)
Agenda

- What is Aggregation?
- RIPE-399 Aggregation Recommendations
- What is happening world wide?
Developed v Developing Internet

- Deaggregation Factor:
  - Routing Table size/Aggregated Size

- Some regions show rampant deaggregation
  - Asia Pacific: 2.48
  - Latin America: 3.40
  - Africa: 2.58

- Compare with:
  - Global Average: 1.85
  - Europe: 1.53
  - North America: 1.69
Observations

- Huge gulf in operational good practices between developing and developed Internet
  - Threatens the very existence of the Internet as we know it
- RIPE-399 is only a recommendation
  - Hopefully all the RIRs will include pointers to it with each address allocation
  - Hopefully more ISPs will pay attention to it
  - Training is there — most ISPs choose to ignore it
Conclusion

- The Internet is in peril as never before
- RIPE-399 now exists
- Make it your BGP good practice document