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?

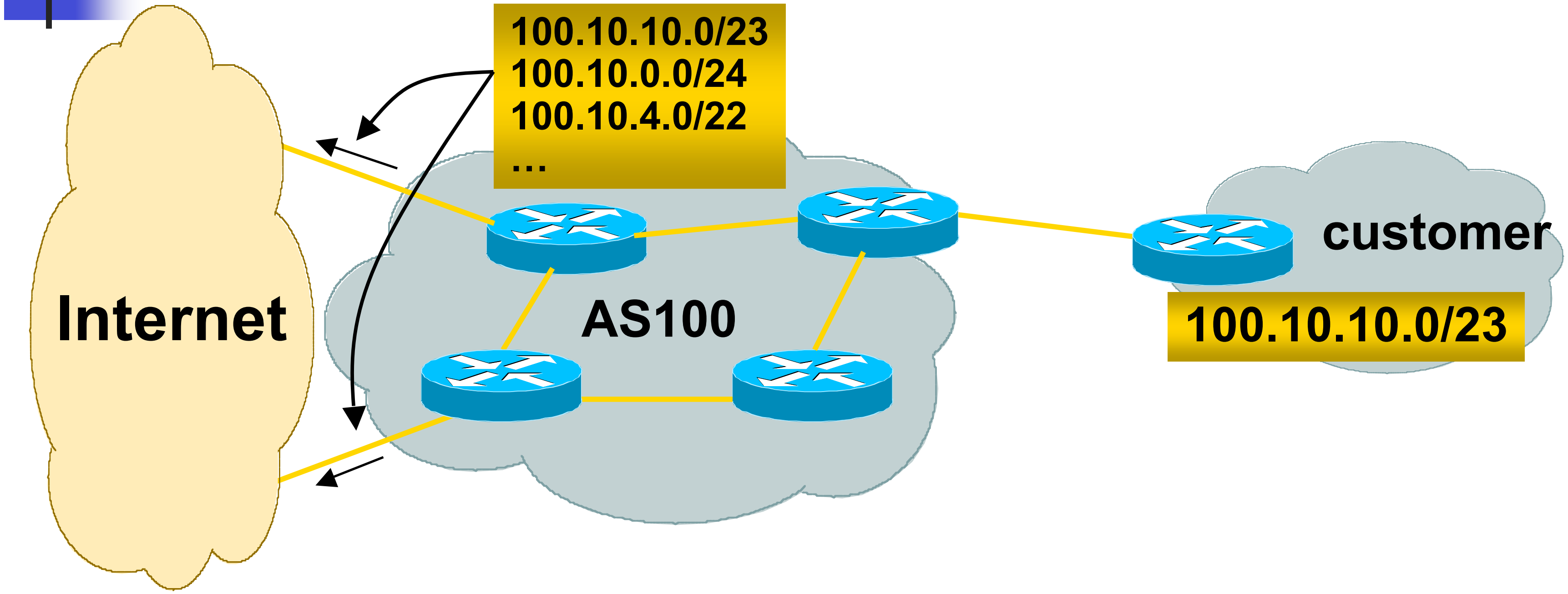
Aggregation

- 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!

Aggregation – Example 1

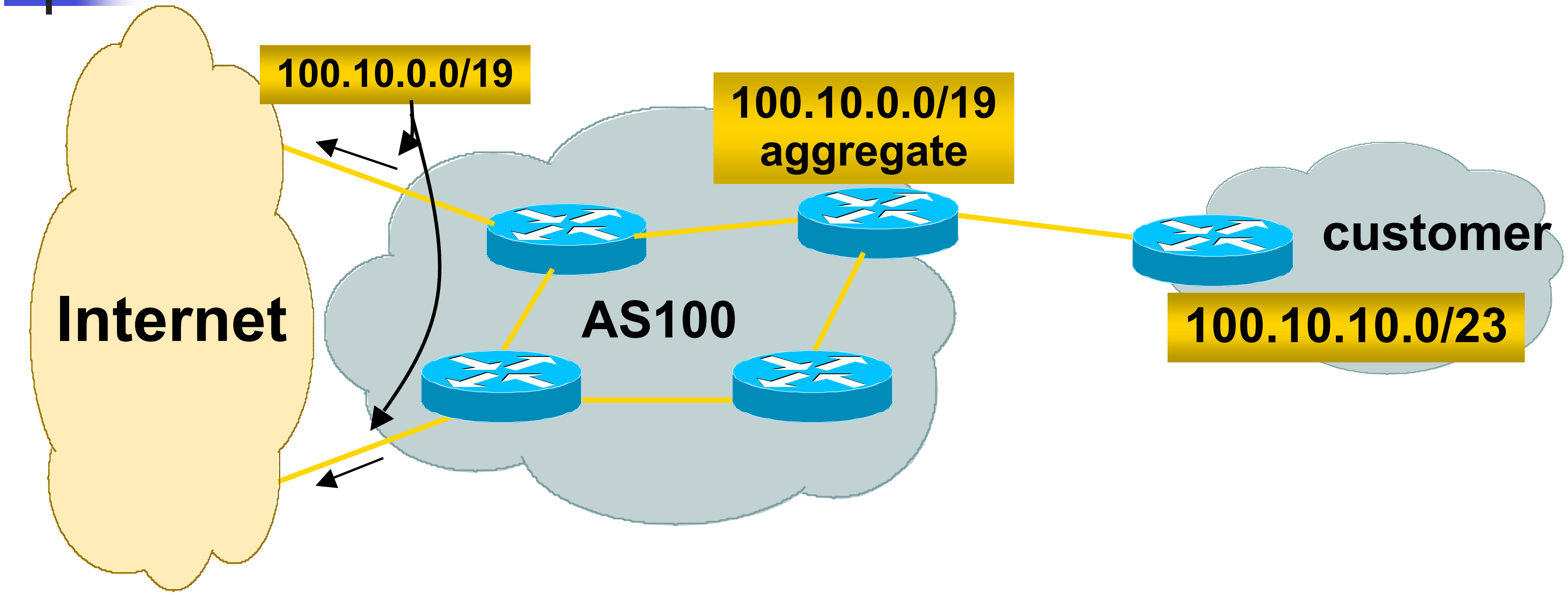


- Customer has /23 network assigned from AS100's /19 address block
- AS100 announces customers' individual networks to the Internet

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  - /23 is withdrawn from AS100’s iBGP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???

Aggregation – Example 2



- 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  - /23 is withdrawn from AS100's iBGP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

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

“The New Swamp”
RIR Space – February 2006

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

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

“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

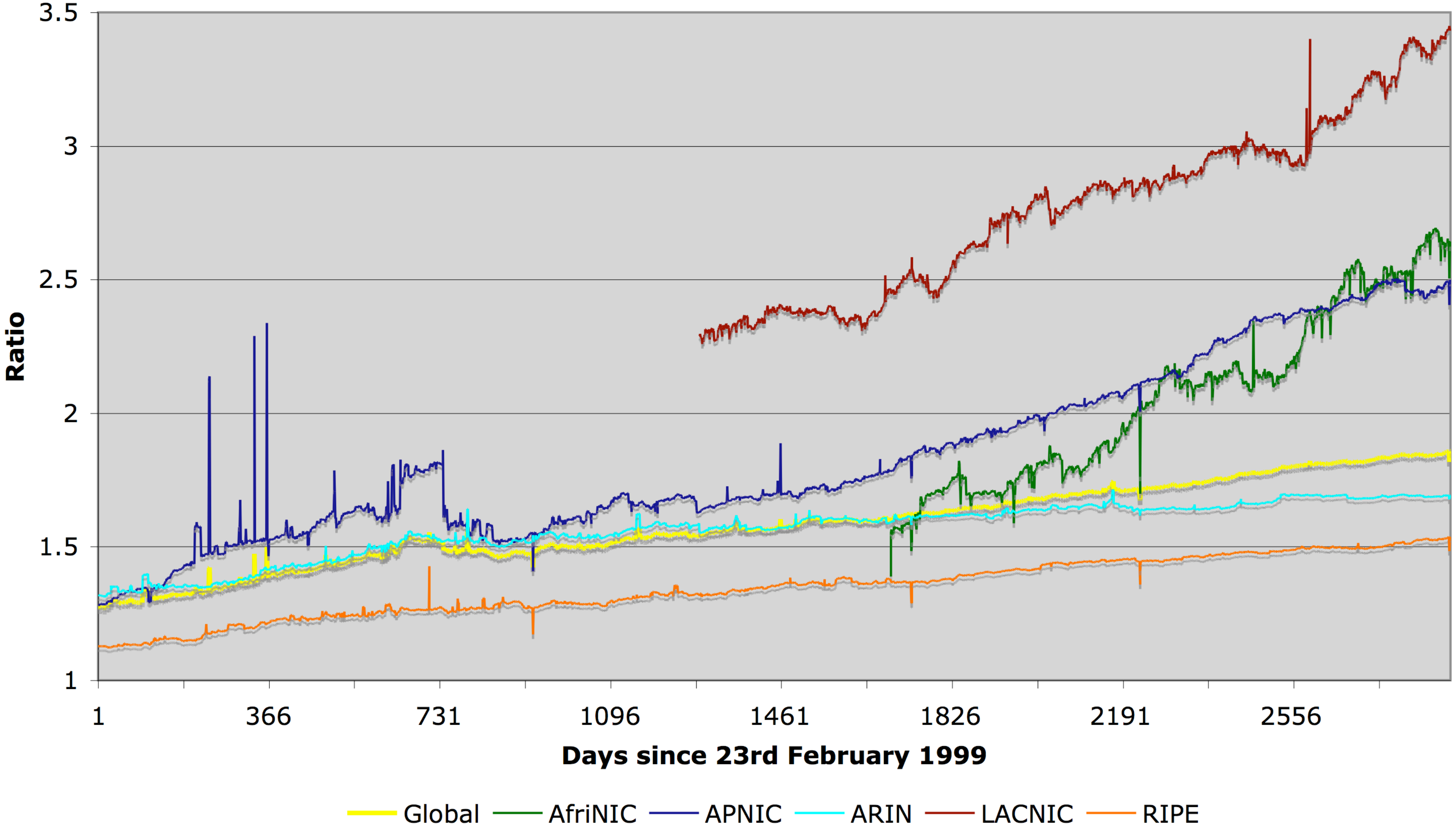
- Deaggregation Factor:
  - Routing Table size/Aggregated Size
- Some regions show rampant deaggregation

Asia Pacific	2.48
Latin America	3.40
Africa	2.58

- Compare with:

Global Average	1.85
Europe	1.53
North America	1.69

Deaggregation: RIR Regions vs Global



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