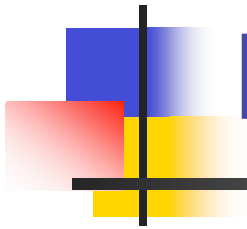
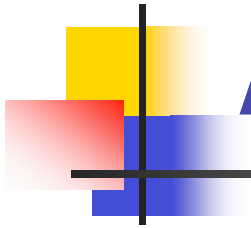


BGP Route Aggregation Best Practices

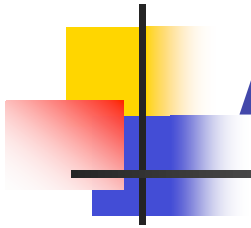
Philip Smith

PhNOG 2
3rd-7th December 2007
Manila, Philippines



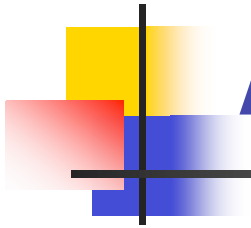
Agenda

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



Aggregation

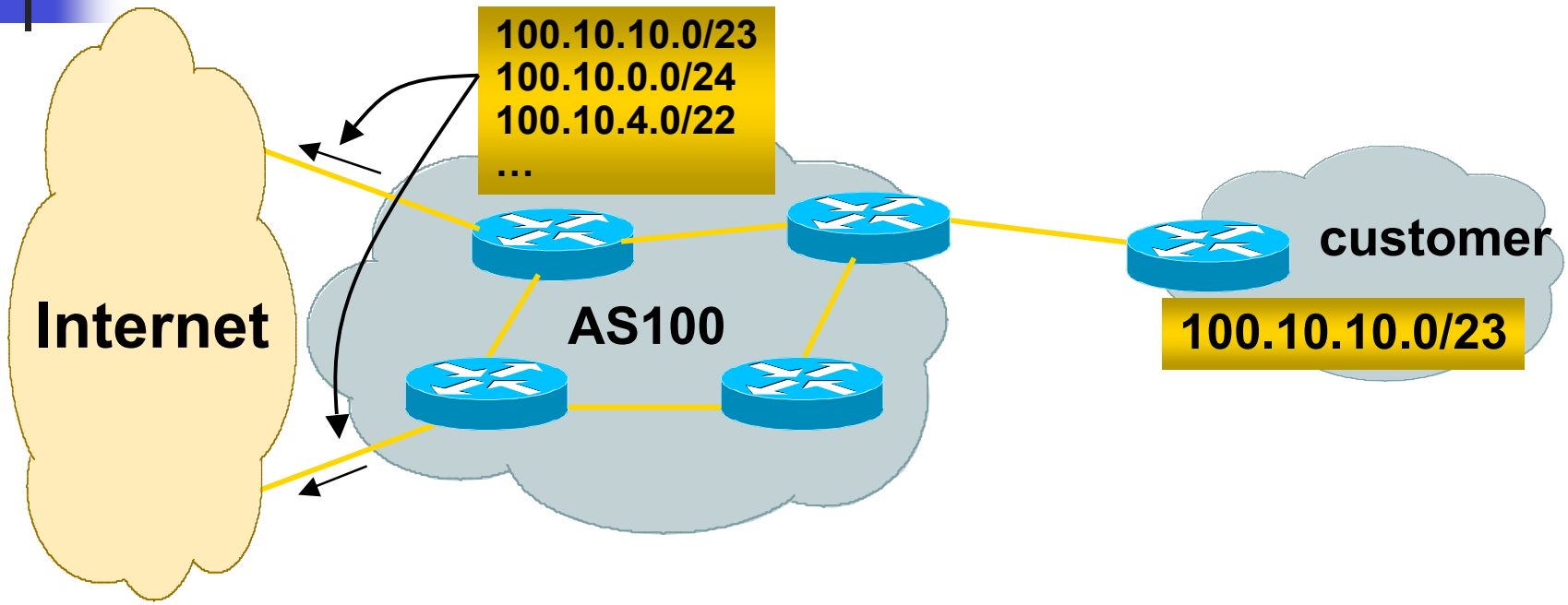
- Aggregation means announcing the address block received from the RIR to the other ASes connected to your network
- Subprefixes of address block must NOT be announced to Internet unless aiding traffic engineering for multihoming
- Subprefixes of this aggregate will be present internally in the ISP network



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 >125000 /24s!

Aggregation – Example 1



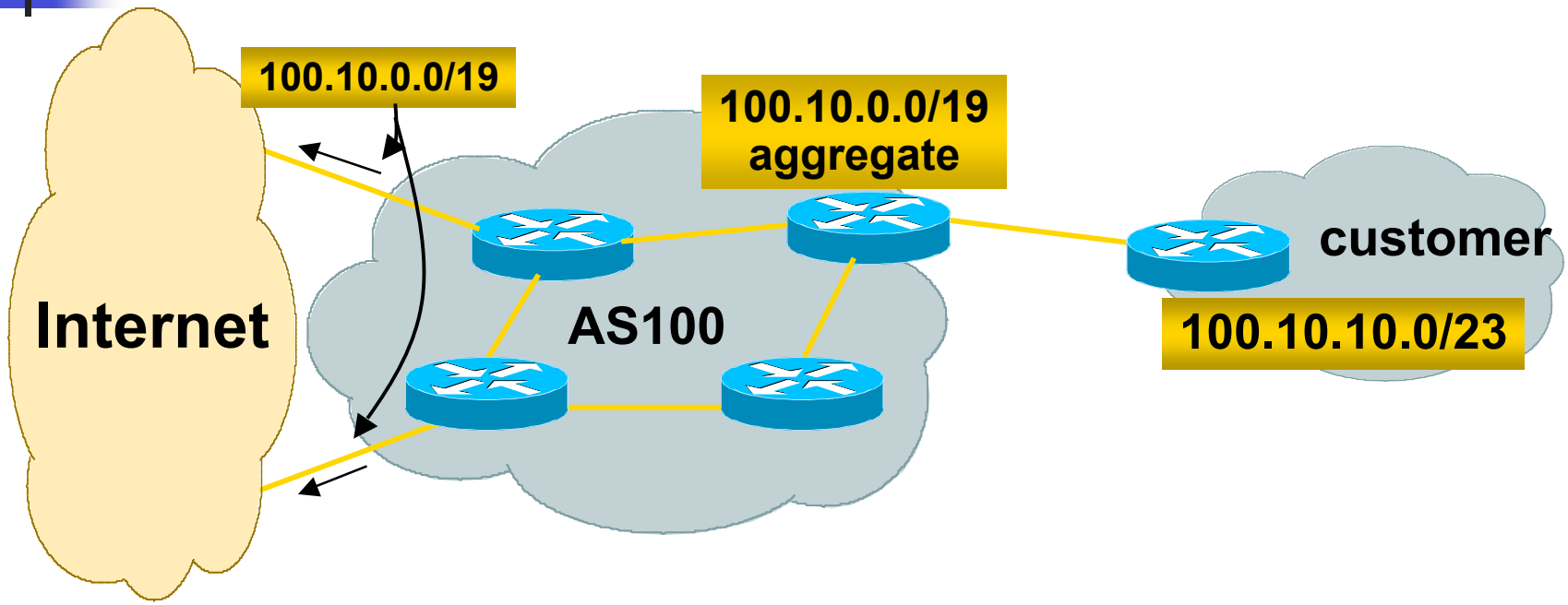
- Customer has /23 network assigned from AS100's /19 address block
- AS100 announces customers' individual networks to the Internet

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

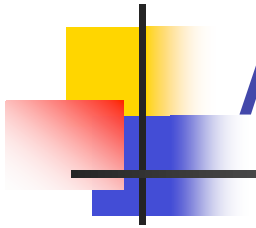
- Customer link returns
  - Their /23 network is now visible to their ISP
  - Their /23 network is re-advertised to peers
  - Starts rippling through Internet
  - Load on Internet backbone routers as network is reinserted into routing table
  - Some ISP's suppress the flaps
  - Internet may take 10-20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

Aggregation – Example 2



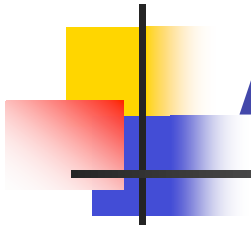
- Customer has /23 network assigned from AS100's /19 address block
- AS100 announced /19 aggregate to the Internet

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

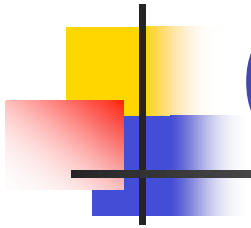
- Customer link returns
- Their /23 network is visible again
  - The /23 is re-injected into AS100's iBGP
- The whole Internet becomes visible immediately
- Customer has Quality of Service perception



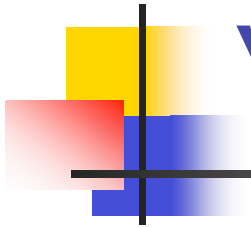
Aggregation – Summary

- 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 (November 2007)



- Current Internet Routing Table Statistics
  - BGP Routing Table Entries 239222
  - Prefixes after maximum aggregation 122992
  - Unique prefixes in Internet 116637
  - Prefixes smaller than registry alloc 111829
  - /24s announced 125530
  - only 5734 /24s are from 192.0.0.0/8
  - ASes in use 26769



“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

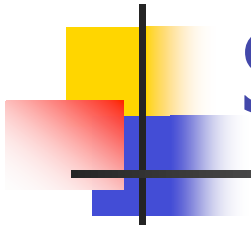
Block	Networks	Block	Networks	Block	Networks	Block	Networks
24/8	**165**	*74/8*	*0*	*124/8*	*0*	**205/8**	**2584**
41/8	*0*	*75/8*	*0*	*125/8*	*0*	**206/8**	**3127**
58/8	*0*	*76/8*	*0*	*126/8*	*0*	**207/8**	**2723**
59/8	*0*	*80/8*	*0*	*188/8*	*0*	**208/8**	**2817**
60/8	*0*	*81/8*	*0*	*189/8*	*0*	**209/8**	**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

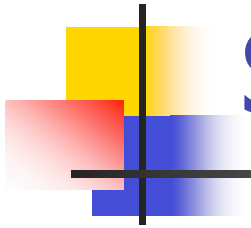
Block	Networks	Block	Networks	Block	Networks	Block	Networks
24/8	3001	74/8	109	124/8	292	205/8	2934
41/8	41	75/8	2	125/8	682	206/8	3879
58/8	606	76/8	4	126/8	27	207/8	4385
59/8	628	80/8	1925	188/8	1	208/8	3239
60/8	468	81/8	1350	189/8	0	209/8	5611
61/8	2396	82/8	1158	190/8	39	210/8	3908
62/8	1860	83/8	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



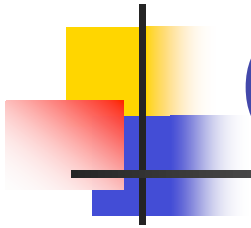
- RIR space shows creeping deaggregation
  - Today an RIR /8 block averages around 6000 prefixes once fully allocated
  - → Existing 74 /8s will eventually cause 444000 prefix announcements
- Food for thought:
  - Remaining 58 unallocated /8s and the 74 RIR /8s combined will cause:
    - 852000 prefixes with 6000 prefixes per /8 density
    - Plus 12% due to “non RIR space deaggregation”
    - → Routing Table size of 954240 prefixes

“The New Swamp” Summary

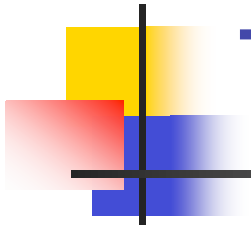


- Rest of address space is showing similar deaggregation too ☹
- What are the reasons?
  - Main justification is traffic engineering
- Real reasons are:
  - Lack of knowledge
  - Laziness
  - Deliberate & knowing actions

BGP Report (bgp.potaroo.net)

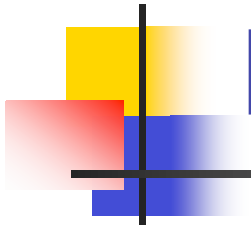


- 199336 total announcements in October 2006
- 129795 prefixes
  - After aggregating including full AS PATH info
    - i.e. including each ASN's traffic engineering
  - 35% saving possible
- 109034 prefixes
  - After aggregating by Origin AS
    - i.e. ignoring each ASN's traffic engineering
  - 10%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



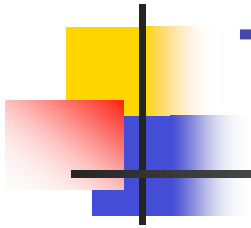
The excuses

- Traffic engineering causes 10% of the Internet Routing table
- Deliberate deaggregation causes 35% of the Internet Routing table



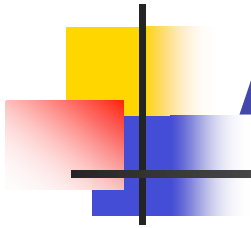
Efforts to improve aggregation

- 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

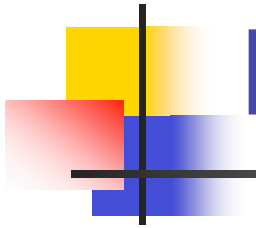
- Also computes the size of the routing table assuming ISPs performed optimal aggregation
- Website allows searches and computations of aggregation to be made on a per AS basis
  - Flexible and powerful tool to aid ISPs
  - Intended to show how greater efficiency in terms of BGP table size can be obtained 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

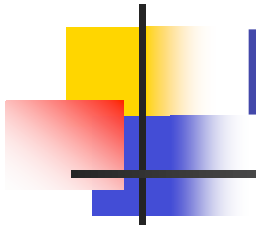
- What is Aggregation?
- RIPE-399 Aggregation Recommendations
- What is happening world wide?

Route Aggregation Recommendations

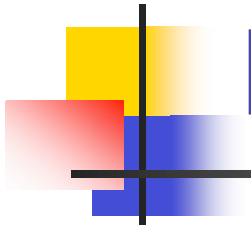


- LINX attempted aggregation policy for members
  - It failed even though most members voted for policy
- RIPE Routing Working Group work item from early 2006
  - Based on early LINX concept
  - Authored by Philip Smith, Mike Hughes (LINX) and Rob Evans (UKERNA)

Route Aggregation Recommendations

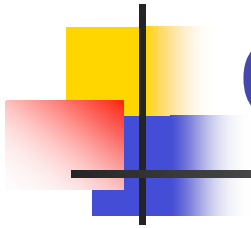


- RIPE Document — RIPE-399
  - http://www.ripe.net/ripe/docs/ripe-399.html
- Discusses:
  - History of aggregation
  - Causes of de-aggregation
  - Impacts on global routing system
  - Available Solutions
  - Recommendations for ISPs



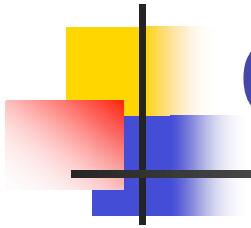
History:

- 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



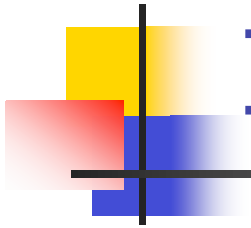
Deaggregation: Claimed causes (1):

- 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"



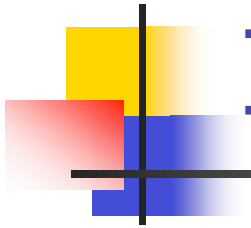
Deaggregation: Claimed causes (2):

- 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 do any **real engineering**
- Legacy Assignments
  - "All those pre-RIR assignments are to blame"
  - In reality it is both RIR and legacy assignments



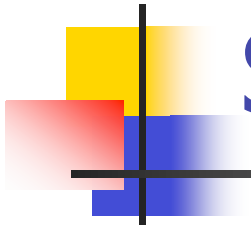
Impacts (1):

- 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



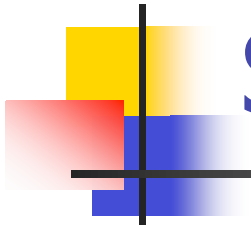
Impacts (2):

- 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



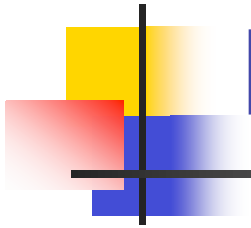
Solutions (1):

- 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”



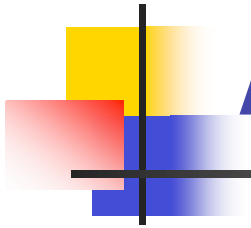
Solutions (2):

- 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



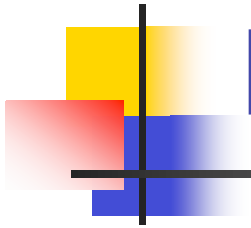
RIPE-399 Recommendations:

- 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)



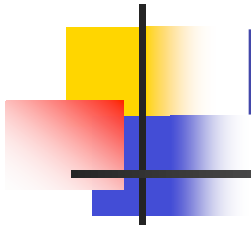
Agenda

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



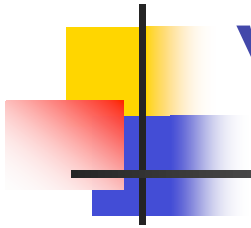
Looking at Deaggregation

- CIDR Report
  - www.cidr-report.org
  - Encourages aggregation following CIDRisation of Internet
  - Today: extensive suite of reports and tools covering state of BGP table
- Routing Report
  - BGP table status on per RIR basis
  - Original CIDR Report and a whole lot more



Deaggregation Factor

- Routing Report
  - One summary takes BGP table and aggregates prefixes by origin AS
    - Called "Max Aggregation" in report
  - Global and per RIR basis
    - http://thyme.apnic.net/current/
- New Deaggregation Factor:
  - Measure of Routing Table size/Aggregated Size
  - Global value has been increasing slowly and steadily since "records began"



“Original Internet” — 2007/11

Total Prefixes

- Global BGP Table
  - 239k prefixes
- North America
  - 112k prefixes
- Europe & Middle East
  - 50k prefixes

Deaggregation Factor

- Global Average
  - 1.94
- North America
  - 1.75
- Europe & Middle East
  - 1.55

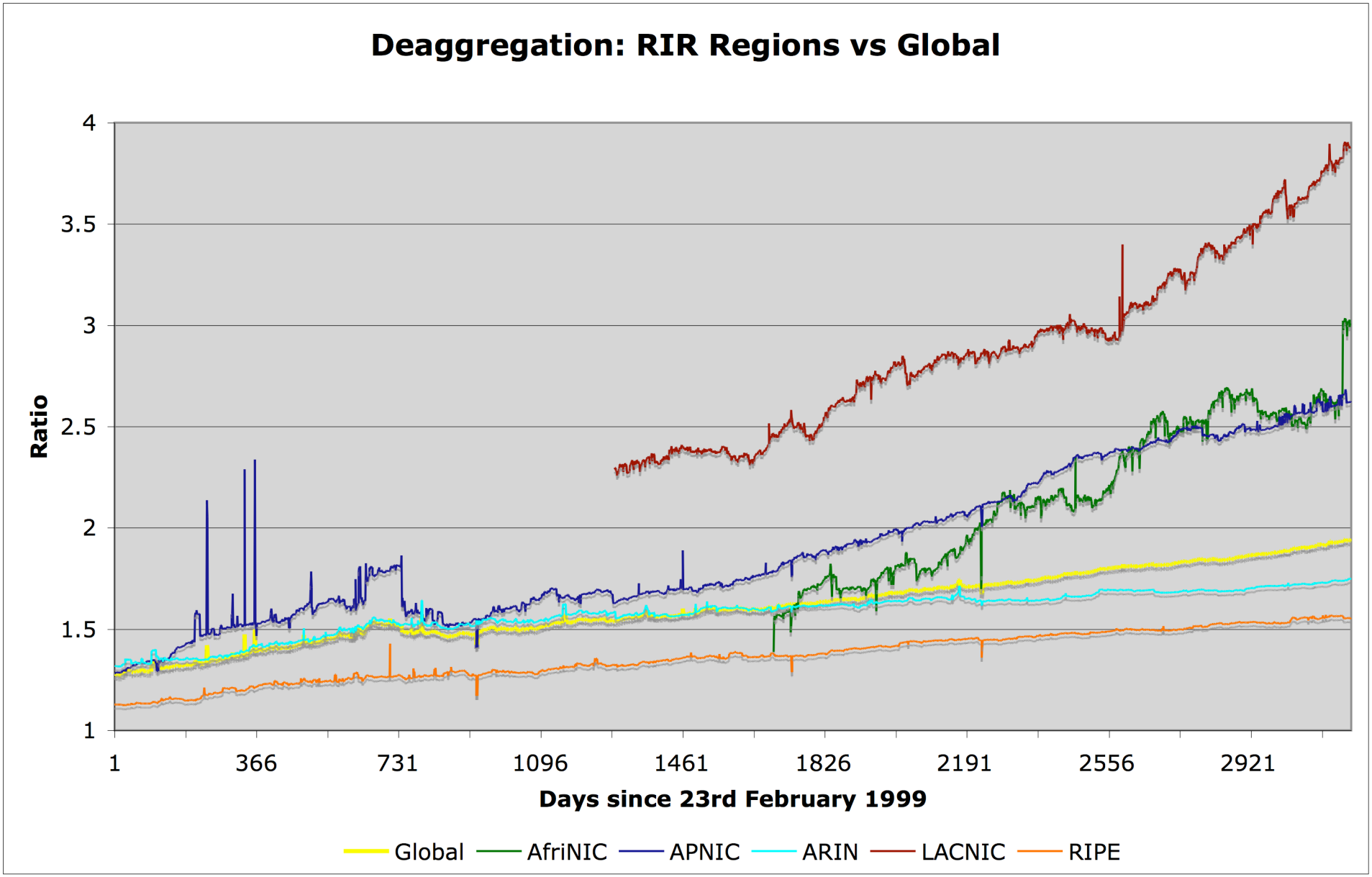
“Newer Internet” — 2007/11

Total Prefixes

- Global BGP Table
  - 239k prefixes
- Asia & Pacific
  - 56k prefixes
- Africa
  - 3k prefixes
- Latin America & Caribbean
  - 18k prefixes

Deaggregation Factor

- Global Average
  - 1.94
- Asia & Pacific
  - 2.62
- Africa
  - 3.01
- Latin America & Caribbean
  - 3.88


Deaggregation: RIR Regions vs Global
Ratio
4
3.5
3
2.5
2
1.5
1
1
366
731
1096
1461
1826
2191
2556
2921
Days since 23rd February 1999
Global
AfriNIC
APNIC
ARIN
LACNIC
RIPE

Africa Aggregation Savings Summary

ASN	No of Nets	Poss Savings	Description
24863	389	352	LINKdotNET AS number
20858	198	195	EgyNet
6713	143	132	Itissalat Al-MAGHRIB
5536	124	114	Internet Egypt Network
33783	123	107	EEPAD TISP TELECOM & INTERNET
33776	100	97	Starcomms Nigeria Limited
24835	88	82	RAYA Telecom - Egypt
15475	81	77	Nile Online
23889	76	61	MAURITIUS TELECOM
20484	64	58	Yalla Online Autonomous Syste
3741	281	55	The Internet Solution
15706	58	54	Sudatel Internet Exchange Aut
29571	55	48	Ci Telecom Autonomous system
12455	36	33	Jambonet Autonomous system
21152	31	26	AS for the uplinks of Soficom
10798	27	26	Standard Bank of South Africa
2018	141	25	Tertiary Education Network
33774	48	22	AS Number for Telecom Algeria
8524	32	22	AUCEGYPT Autonomous System
15804	22	21	AS of The Way Out Internet So

http://thyme.apnic.net/current/data-CIDRnet-AFRINIC

Asia & Pacific Aggregation Savings Summary

ASN	No of Nets	Poss Savings	Description
4755	1464	1395	Videsh Sanchar Nigam Ltd. Aut
9498	1051	987	BHARTI BT INTERNET LTD.
17488	879	809	Hathway IP Over Cable Interne
4134	1103	800	CHINANET-BACKBONE
9583	1090	673	Sify Limited
7545	720	598	TPG Internet Pty Ltd
18101	611	555	Reliance Infocom Ltd Internet
4668	520	510	LG-EDS Systems Inc.
9829	519	509	BSNL National Internet Backbo
4766	816	487	Korea Telecom (KIX)
4812	524	447	China Telecom (Shanghai)
17676	503	439	Softbank BB Corp.
17974	398	384	PT TELEKOMUNIKASI INDONESIA
4808	504	382	CNCGROUP IP network: China169
9443	440	370	Primus Telecommunications
4802	481	340	Wantree Development
4538	355	320	China Education and Research
7552	296	292	Vietel Corporation
9929	325	276	China Netcom Corp.
4780	308	264	Digital United Inc.

http://thyme.apnic.net/current/data-CIDRnet-APNIC

North America Aggregation Savings Summary

ASN	No of Nets	Poss Savings	Description
11492	1173	1131	Cable One
18566	1032	1023	Covad Communications
4323	1377	1014	Time Warner Telecom
6478	1124	945	AT&T Worldnet Services
22773	819	765	Cox Communications, Inc.
5668	664	647	CenturyTel Internet Holdings,
19262	812	629	Verizon Global Networks
15270	603	562	PaeTec.net -a division of Pae
6517	584	551	Yipes Communications, Inc.
19916	569	522	OLM LLC
6197	1030	521	BellSouth Network Solutions,
2386	1271	490	AT&T Data Communications Serv
855	546	484	Canadian Research Network
33588	470	447	Bresnan Communications, LLC.
3356	833	418	Level 3 Communications, LLC
7011	982	412	Citizens Utilities
20115	846	393	Charter Communications
11139	380	352	Cable & Wireless Dominica
3464	371	346	Alabama SuperComputer Network
3602	414	338	Sprint Canada, Inc.

http://thyme.apnic.net/current/data-CIDRnet-ARIN

Latin America Aggregation Savings Summary

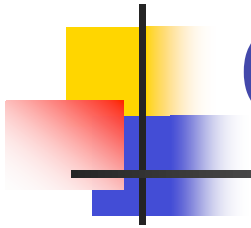
ASN	No of Nets	Poss Savings	Description
8151	1161	946	UniNet S.A. de C.V.
11830	539	530	Instituto Costarricense de El
16814	427	417	NSS, S.A.
14522	381	373	SatNet S.A.
7303	380	321	Telecom Argentina Stet-France
6471	357	319	ENTEL CHILE S.A.
11172	375	315	Servicios Alestra S.A de C.V
22047	317	304	VTR PUNTO NET S.A.
10481	301	291	Prima S.A.
14117	285	270	Telefonica del Sur S.A.
11556	236	232	Cable-Wireless Panama
6147	242	222	Telefonica Del Peru
10620	260	217	TVCABLE BOGOTA
28573	213	186	NET Servicos de Comunicao S.A
7738	206	182	Telecomunicacoes da Bahia S.A
23216	220	173	RAMtelecom Telecomunicaciones
20299	199	169	NEWCOM AMERICAS
19169	183	164	Telconet
8163	169	154	METROTEL REDES S.A.
21826	186	139	INTERCABLE

http://thyme.apnic.net/current/data-CIDRnet-LACNIC

EU & Middle East Aggregation Savings Summary

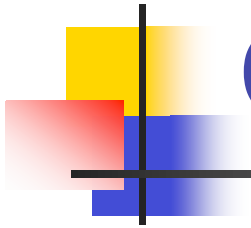
ASN	No of Nets	Poss Savings	Description
9116	347	319	Goldenlines main autonomous s
8551	327	287	Bezeq International
8866	280	254	Bulgarian Telecommunication C
8452	239	232	TEDATA
3352	235	194	Ibernet, Internet Access Netw
12479	195	189	Uni2 Autonomous System
9121	210	185	TTnet Autonomous System
3215	264	173	France Telecom Transpac
3269	230	158	TELECOM ITALIA
5462	174	149	Telewest Broadband
5486	154	136	Euronet Digital Communication
6830	170	133	UPC Distribution Services
29357	135	131	WATANIYA TELECOM
9155	129	119	QualityNet AS number
15471	180	116	SNR - Societatea Nationala de
3300	214	112	AUCS Communications Services
31083	105	98	PowerNet.BG, Sofia, Bulgaria
9051	149	94	INCONET Autonomous System
15611	96	94	Iranian Research Organisation
12883	97	93	Farlep-Internet ISP

http://thyme.apnic.net/current/data-CIDRnet-RIPE



Observations

- Huge gulf in operational good practices between "older" and "newer" Internet
  - Could threaten the Internet as we know it
- RIPE-399 is only a recommendation
  - Hopefully all the RIRs will include pointers with each address allocation
  - Hopefully more ISPs will pay attention to it
  - Training is there — most ISPs choose to ignore it



Conclusion

- "Newer" Internet is growing rapidly
  - As is the deaggregation there
- RIPE-399 now exists
- Make it your BGP good practice document