UA-IX

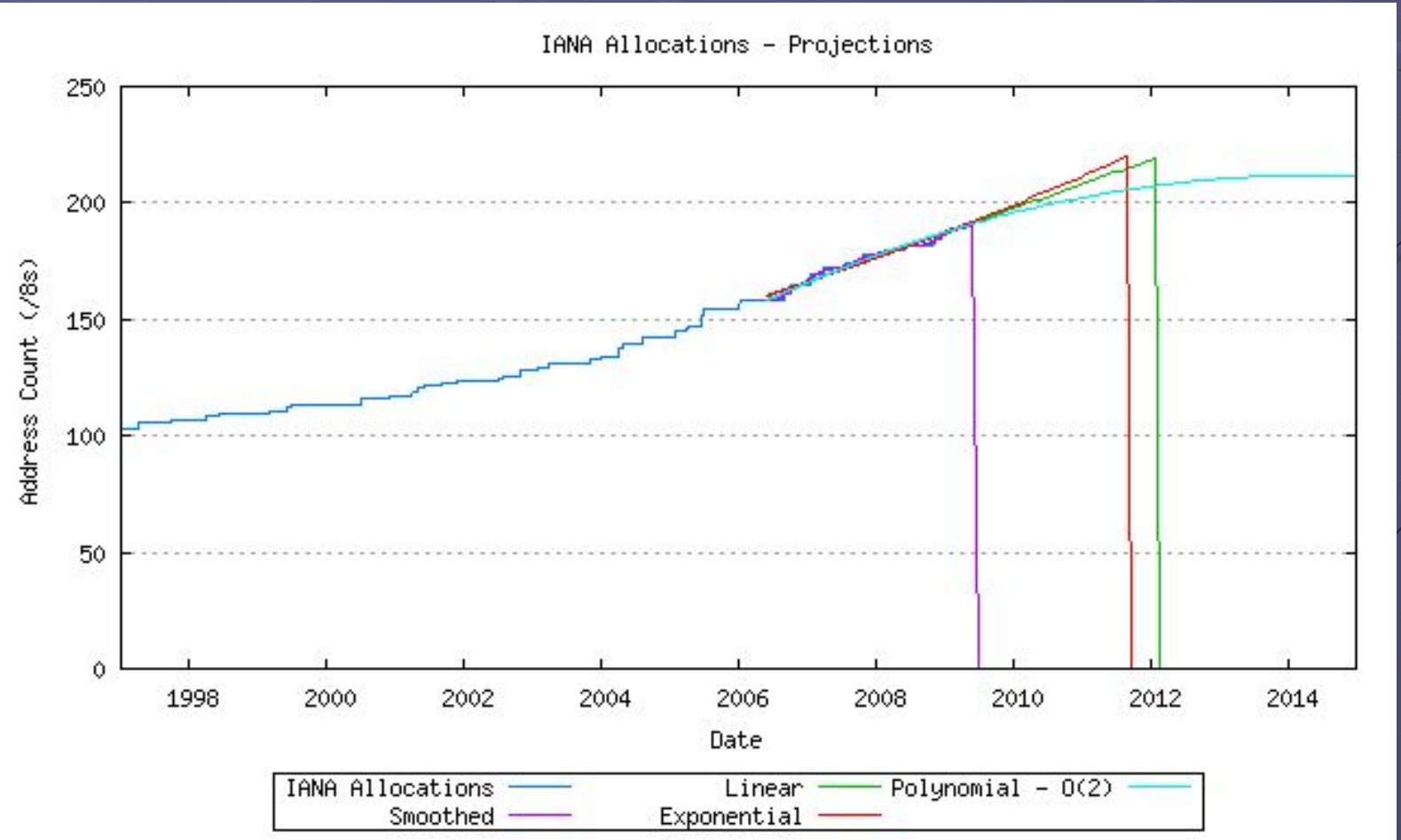
Новые технологии и возможности Сети

Сергей Полищук

29 мая 2009 г.

IPv6

Пул 32bit IPv4 адресов состоит из 220 /8 (36 reserved by IETF)
IANA -> RIR -> LIR -> End User
190 из 220/8 уже выданы.
http://www.potaroo.net/tools/ipv4/
Ожидаемая дата исчерпания адресного пула IANA:
13 Июня 2011



27.11.2008 Правление ИнАУ приняло решение:
3.4. провести тестування впровадження ipv6 unicast з Учасниками Мережі на добровільній основі.

01.04.2009 на базе RS-I началось тестирование.

Четверо Участников сделали это:

Топнет, Датагруп, Нетассист, УНТ

Для участия в эксперименте нужно лишь сообщить по адресу staff@ix.net.ua следующие параметры:
- номер своей автономной системы
- список анонсируемых префиксов ipv6
- мак-адрес маршрутизатора и идентификатор своего порта включения в UA-IX (для тех у кого более одного порта включения)

```
BGP router identifier 195.35.65.1, local AS number 15645
BGP table version is 49, main routing table version 49
4 network entries using 624 bytes of memory
4 path entries using 304 bytes of memory
3779/4 BGP path/bestpath attribute entries using 634872 bytes of memory
3438 BGP AS-PATH entries using 129400 bytes of memory
164 BGP community entries using 4408 bytes of memory
0 BGP route-map cache entries using 0 bytes of memory
0 BGP filter-list cache entries using 0 bytes of memory
Bitfield cache entries: current 4 (at peak 5) using 124 bytes of memory
BGP using 769732 total bytes of memory
Dampening enabled. 0 history paths, 0 dampened paths
BGP activity 6941/3705 prefixes, 38762/32815 paths, scan interval 60 secs
Neighbor        V           AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
2A02:280:0:FFFF::34
                4      21011   89546   80970       49    0    0 1w2d            1
2A02:280:0:FFFF::58
                4      30955   64154   63827       49    0    0 1w2d            1
2A02:280:0:FFFF::90
                4      29632  159438  155937       49    0    0 1w2d            1
2A02:280:0:FFFF::225
                4      21219  167075  158678       49    0    0 1w2d            1
BGP table version is 49, local router ID is 195.35.65.1
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
              r RIB-failure, S Stale
Origin codes: i - IGP, e - EGP, ? - incomplete
   Network          Next Hop            Metric LocPrf Weight Path
*> 2A01:D0::/32     2A02:280:0:FFFF::90
                                                           0 29632 i
*> 2A01:758::/32    2A02:280:0:FFFF::34
                                                           0 21011 i
*> 2A02:70::/32     2A02:280:0:FFFF::225
                                                           0 21219 i
*> 2A02:CF0::/32    2A02:280:0:FFFF::58
                                         0                 0 30955 i
```

В чем отличие?

Multicast neighbor discovery вместо ARP,
Multicast router advertisement (RA) вместо default route

Другой синтаксис написания адресов.
2001:db8**::**/32 =
2001:0db8:0000:0000:0000:0000:0000:0000/32

Каждый интерфейс уже имеет **link local address**
FE80::<64 bit MAC>/10

Формат глобального IPv6 адреса:
<3 bit “001”><45bit global prefix><16 bit SLA><64 bit MAC>

Казалось бы, имеем 128 бит вместо 32, но реально…

LIR получает блок /32, префикс которого занимает 28bit:
0010 0000 0000 0001:1101 1010 1000: :/32

Начиная с /64 начинает работать механизм Stateless Autoconfiguration. Клиентское устройство самостоятельно присваивает себе глобальный адрес, приписав к изученному RA свой мак-адрес, автоматически лишая нас возможности использования 64 бит для указания фиксированного адреса. Еще 19 бит уходят на «001» и SLA, остается лишь 48… Отнимаем еще 28 бит занимаемых собственно префиксом и получаем, что LIR может распоряжатся лишь 20 битами адреса, т.е. владеет 1 миллионом адресов (как IPv4 /12)

Особенности передачи данных на скоростях >1GE

FLOW Control

В 1997-м году сервера были недостаточно быстрыми и тогда, при разработке стандарта GigabitEthernet в спецификацию был добавлен механизм приостановки передачи трафика **flow-control** (IEEE 802.3x).

В случае, если принимающая сторона не способна принимать трафик, она отправляет фрейм на специальный малтикастовый адрес: 01:80:c2:00:00:01

В отличие от протокола TCP, данная функциональность работает на уровне одного линка (между двумя интерфейсами) и не зная информации о состоянии всей сети, блокирует весь трафик на интерфейсе, дублируя при этом более эффективный механизм протокола TCP, умеющий управлять скоростью передачи отдельных сессий в зависимости от множества параметров актуальных на участке между любыми отправителями и получателями данных в глобальных сетях.

Чаще всего проблемы возникают в коммутаторах работающих одновременно на нескольких скоростях. Например, в UA-IX для подключения Участников доступны скорости 100/1000/10000. Гипотетически, при попытке передачи информации на максимальной скорости из порта 10000 в порт 1000, должен срабатывать механизм ethernet flow control.

На практике, ввиду неоднозначности понимания данного механизма разными производителями, последствия срабатывания flow control непредсказуемы. Как правило блокируется прохождение легитимного трафика, а не только того, который привел к перегрузке. Чаще всего срабатывание механизма выглядит как понижение реальной пропускной способности всех портов до уровня скорости самого медленного порта коммутатора. Например, в такой ситуации на всех портах коммутатора Foundry BigIron будет выставлен статус блокировки всего входящего трафика. Дешевые неуправляемые коммутаторы без поддержки протокола IEEE 802.3x таят в себе еще больше опасности, т.к. Pause Frame будет ретранслирован во все порты такого коммутатора как и положено для малтикаста, что приведен к блокировке всех портов устройств подключенных ко всей цепочке таких коммутаторов.

Как правило, современные коммутаторы известных производителей обеспечивают передачу всего трафика на полной скорости порта, что минимизирует вероятность срабатывания flow-control. Большинство Участников UA-IX работают именно с таким оборудованием, тем не менее, за год мы наблюдали около 3-4 случаев "залипания« 10GE портов вызванных феноменом flow-control.

Текущая статистика портов Участников которые активировали у себя flow-control:

Port	Flow Cntrl	Port	Flow Cntrl
DG	ASYM	ETT	SYM
Kancom	SY/ASY	Ukrsat	SYM
volz	ASYM	Navigato>	ASYM
LuckyNet	SYM	Techsyst>	SY/ASY
Citius	SY/ASY	Navigato>	ASYM
United	ASYM	VikaTV	ASYM
UMC	SY/ASY	Wimax	SY/ASY
Astelit	SY/ASY	Kumirtel>	SYM
Lviv.Net	SY/ASY	Cosmonov>	SY/ASY
Merlin	SY/ASY	Mobicom	SYM
Adamant	SYM	Intertel	SYM
Tenet	SY/ASY		

Хорошая новость. В версии XOS 12.1.3 работающей сейчас на BD8810 появилась возможность блокировать прохождение фреймов flow-control в обоих направлениях и теперь сеть UA-IX защищена даже от единичных непредсказуемых срабатываний.

Вывод: крайне НЕ РЕКОМЕНДУЕТСЯ отрабатывать Pause Frames на уровне ядра сети. Наиболее верное решение – распознавать такие фреймы и игнорировать их.

JUMBO

Jumbo - так звали гигантского слона из Африки (1861-1885) прославившегося в Париже и Лондоне.

Но какое отношение имеют слоны к сети UA-IX
и почему мы об этом говорим?

С момента своего появления (~1980) размер Ethernet-фреймов всегда был 1500.

Jumbo frame - это любой Ethernet frame размеры полезной нагрузки которого превышают обычные 1500 байт. Данное понятие доступно только для сетей работающий на скоростях 1000 и выше.

Интересно, что данное понятие не вошло в стандарт IEEE 802, и каждый из производителей выбирает свой максимальный размер фреймов доступных на его оборудовании, но все же федеральные сети США совместно с проектом **Internet2** популяризовали значение MTU **9000** для Jumbo-фреймов на своих сетях, хотя теоретически фрейм можно было бы растянуть до 64000 байт (ограничение IPv4).

Почему 9000? Причины две: алгоритм 32 bit CRC теряет свою эффективность при размеру >12000, 8Кб фрейм протокола NFS помещается в 9000.

Коммутаторы Extreme Networks, а также сеть UA-IX поддерживают размер 9216 байт.

Зачем это нужно?

При прохождении каждого пакета по сети передачи данных, на каждом из активных устройств анализируется заголовок в котором указана информация об отправителе и получателе, а затем, принимается решение о том, куда пойдет этот пакет дальше. Использование jumbo frames означает, что через сеть будет передаваться намного меньше пакетов, что приведен к снижению нагрузки как на процессор, так и увеличит полезную пропускную способность каналов. Один фрейм 9000 заменяет 6 обычных 1500-байтных, уменьшая нагрузку на сеть в пять раз, а также сокращая объем передаваемой служебной информации на 290=((5*(40+18)) байт для каждого TCP/IP пакета.

1GE не способен пропустить более 83000 стандартных пакетов – он будет полностью забит, заставляя процессор или сетевую Карту принимать решение о передаче данных 83000 раз в секунду. В случае с 9К-пакетами получим всего 13900pps и дополнительно 32Mbps реальной полосы освобожденной от оверхеда.

Проблемы?

Маркетологи производителей процессоров и маршрутизаторов против.

Эффективно только если в цепочке устройств все используют Jumbo.

Если слон попытается залезть в бутылочное горлышко, требуется Механизм его гарантированного прохождения.

К сожалению, описанный в RFC 1191 & 1981 path MTU discovery не всегда работает из-за фильтрации ICMP некоторыми провайдерами – требуется поддержка альтернативного механизма **RFC 4821 robust path MTU discovery.**

Решив эти проблемы, возможно смешение сетей с произвольными MTU, что позволит значительно снизить нагрузку на сетевое оборудование.

Одна из причин задержки появления стандарта 100GE заключается в попытке сохранить
старый малый размер MTU, но наталкивается на физические ограничения, т.к. на обработку одного фрейма требуется потратить не более 0.12 uS, т.е. от интерфейса
ожидают производит ельность в 1000 более высокую, чем для FastEthernet, но технологические процессы производства микрокристаллов за 15 лет подняли производительность лишь в сотни раз.

Переход на размер пакета в 64К позволил бы запустить 100GE на той же дешевой элементной базе, которая сейчас используется для 1GE. IPv6 дает возможность в будущем еще больше увеличивать размер MTU:

Текущие параметры				**Альтернатива v6**		**Альтернатива v4**	
Speed	**Year**	**MTU**	**Wire Time**	**MTU**	**Wire Time**	**MTU**	**Wire Time**
10	1982	1.5 kB	1200 uS				
100	1995	1.5 kB	120 uS	9 kB	720 uS	4.3 kB	433 uS
1000	1998	1.5 kB	12 uS	64 kB	512 uS	9 kB	72 uS
10000	2002	1.5 kB	1.2 uS	150 kB	120 uS	64 kB	51.2 uS
100000			0.12 uS	1.5 MB	120 uS	64 kB	5.12 uS
1000000				15 MB	120 uS	64 kB	0.512 uS

Эффективное подключение на скорости 10GE к UA-IX

20 января 2009 17:12:26 – с этого момента, в сети UA-IX появился новый центральный коммутатор **Extreme Networks Summit X650**, целевое назначение которого - подключение оборудования UA-IX и Участников интерфейсами 10GE.

В состав устройства входит 24 порта 10GE формфактора **SFP+** Дальнейшее расширение возможно путем каскадирования X650 до 176 портов 10GE. Таким образом, после установки первого коммутатора X650, потенциальная емкость сети UA-IX выросла с 536 до 780Gbps с возможностью дальнейшего расширения до 2296Gbps.

С этого момента, при подключении на скорости 10GE, на технической площадке DG, от Участника будет требоваться предоставление оптического модуля формфактора SFP+

Поскольку на уровне оптического сигнала обеспечивается interoperability, на оборудовании самого Участника, кроме нового SFP+ может использоваться также любой из форматов 10GBASE XENPAK, 10GBASE X2, 10GBASE XFP.

Справочная сравнительная информация по ценам на оптические модули:

Part-N	GPL	Name Description
10301	$1255	10GBASE-SR SFP+ up to 300 meters Multimode Fiber
10302	$2096	10GBASE-LR SFP+ up to 10 km Singlemode Fiber
04	$126	10GBASE-CR SFP+ 1m pre-terminated twin-ax copper cable
10305	$173	10GBASE-CR SFP+ 3m pre-terminated twin-ax copper cable
06	$210	10GBASE-CR SFP+ 5m pre-terminated twin-ax copper cable
07	$252	10GBASE-CR SFP+ 10m

«Старые» модули формата XFP несколько дороже:
10121 $2096 SR XFP 10GBASE-SR XFP up to 300 meters Multimode Fiber
10122 $3147 LR XFP 10GBASE-LR XFP up to 10 km Singlemode Fiber

Те участники, оборудование которых находится на расстоянии менее 10 метров до стойки 1-2 на площадке DG, могут значительно сэкономить путем использования специальных пассивных патчкордов стандарта 10GBASE-CR.

В случае использования такого патчкорда, полностью отпадает необходимость в дорогостоящих оптических модулях и стоимость организации канала связи 10GE фактически снижается более, чем в 10 раз.

Технология SFP+ поддерживается ведущими производителями:

Extreme Networks, Cisco Systems (только новые серии коммутаторов), **Force 10**

http://www.cisco.com/en/US/prod/collateral/modules/ps5455/data_sheet_c78-455693.html
http://www.force10networks.com/products/mediaspecifications.asp

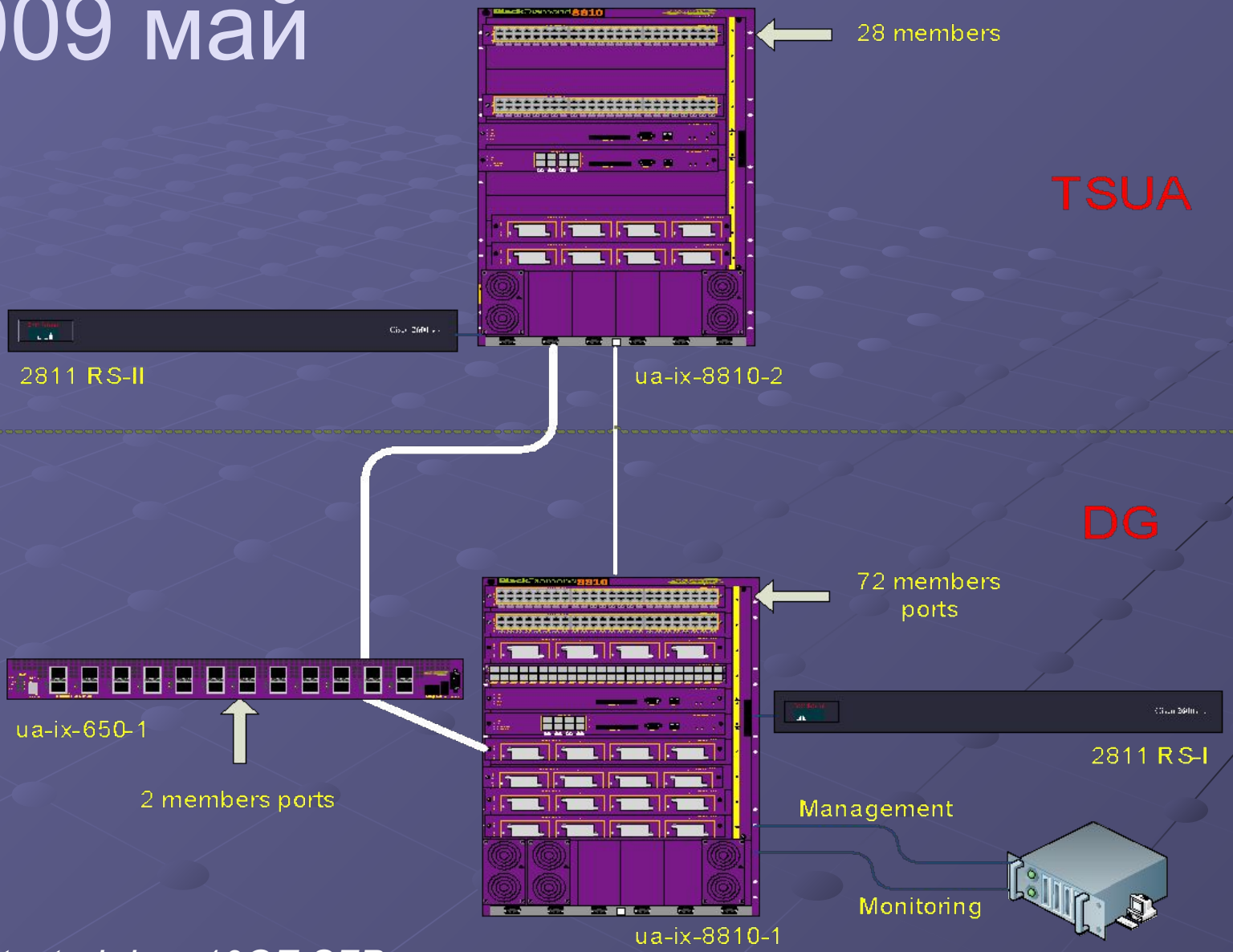
Extreme Networks для своих коммутаторов Summit X450a/e и X350 производит новый модуль расширения XGM2-2sf на два порта SFP+.

При желании, любой из 15 участников уже подключенных к ua-ix-8810-1 модулем XFP, может быть переключен на новый коммутатор ua-ix-650-1 путем простой замены оптических модулей.

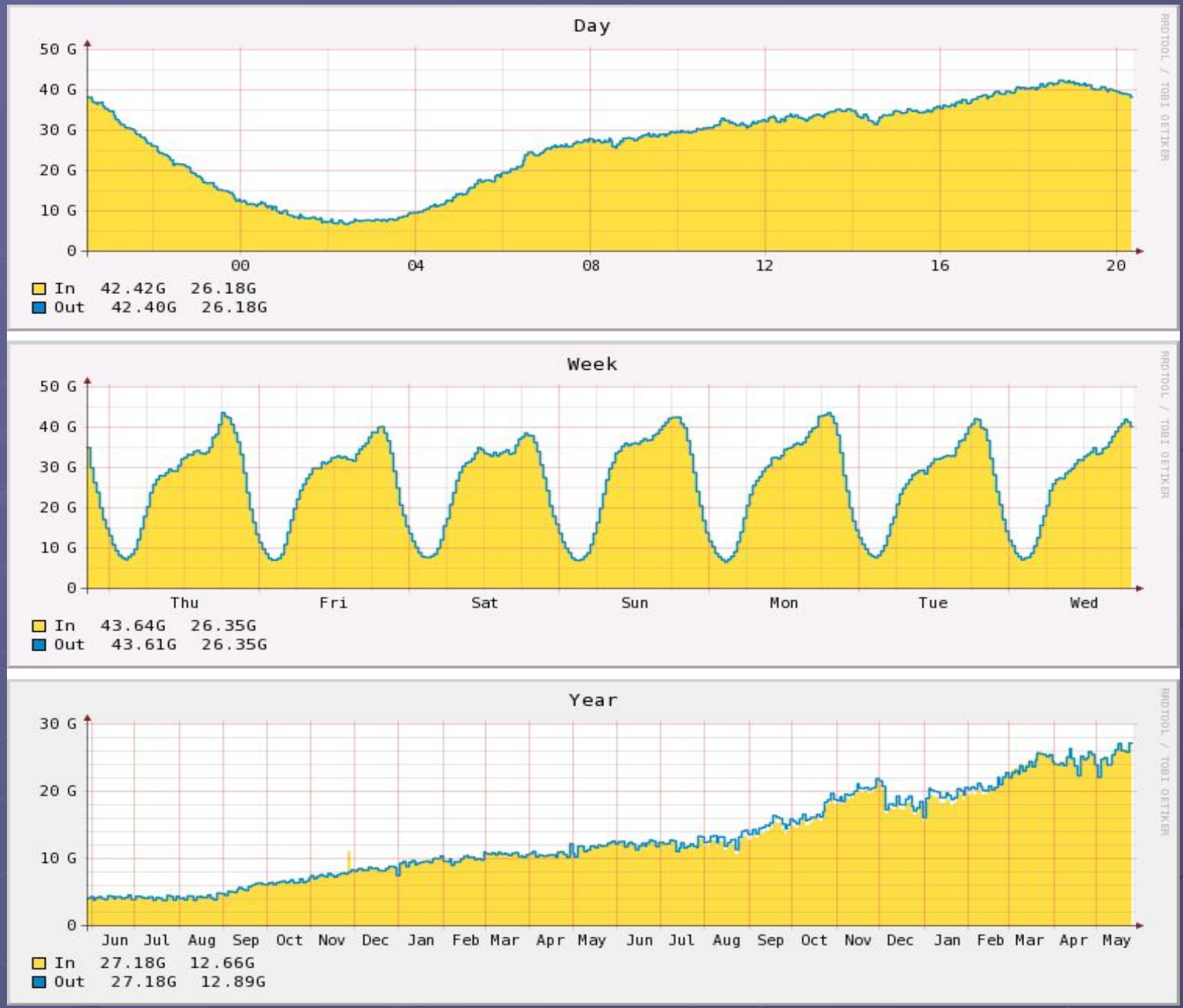
С одной стороны, это позволит Участнику использовать более дорогой модуль для развития собственной сети, а с другой – возможность дальнейшего расширения скорости своего подключения до 80GE (с шагом 10GE).

UA-IX сегодня

2009 май

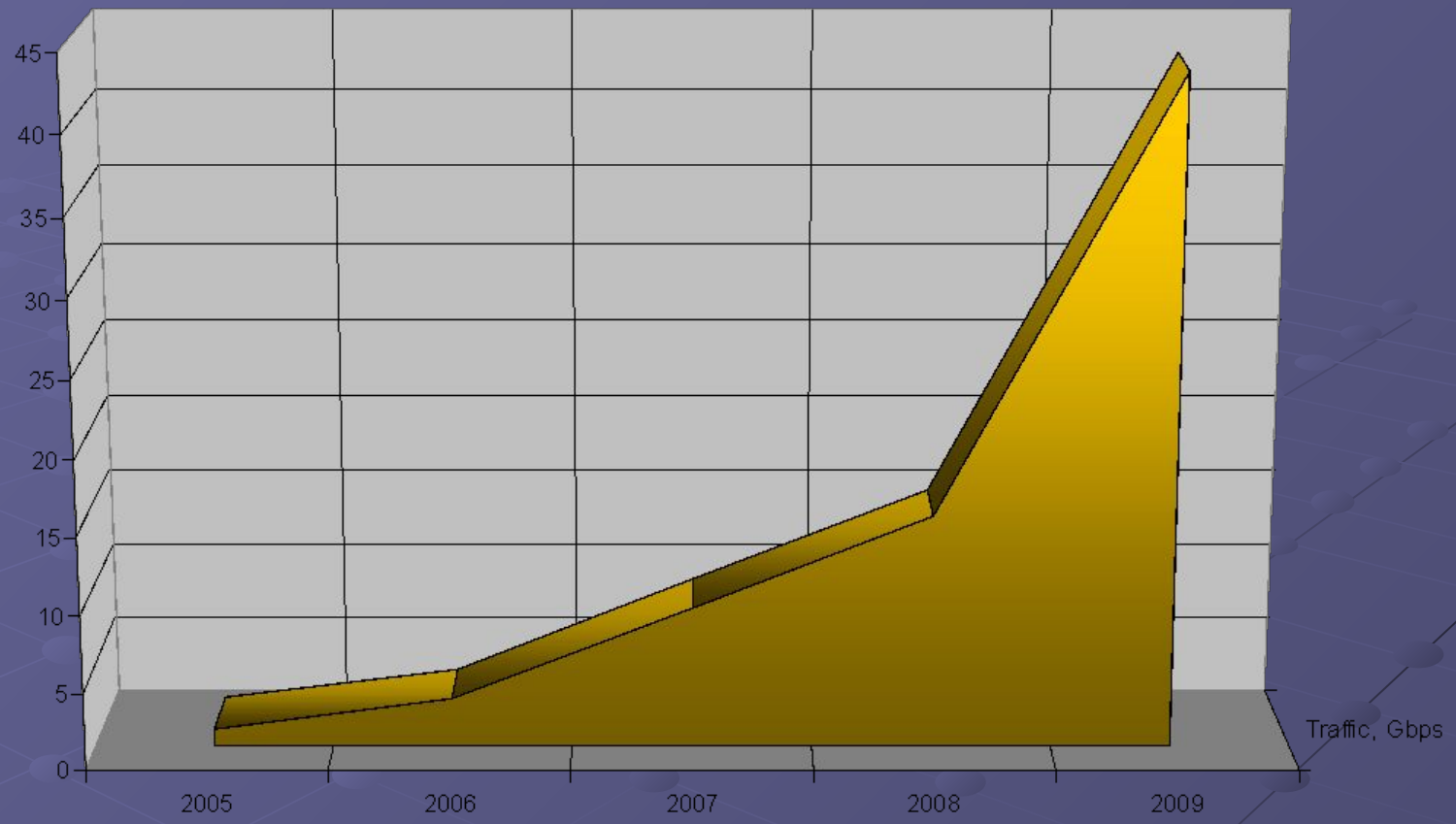


EAPS protected ring, 10GE SFP+


Day
50 G
40 G
30 G
20 G
10 G
0
00
04
08
12
16
20
In 42.42G 26.18G
Out 42.40G 26.18G
RRDTOOL / TOBI OETIKER
Week
50 G
40 G
30 G
20 G
10 G
0
Thu
Fri
Sat
Sun
Mon
Tue
Wed
In 43.64G 26.35G
Out 43.61G 26.35G
RRDTOOL / TOBI OETIKER
Year
30 G
20 G
10 G
0
Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May
In 27.18G 12.66G
Out 27.18G 12.89G
RRDTOOL / TOBI OETIKER

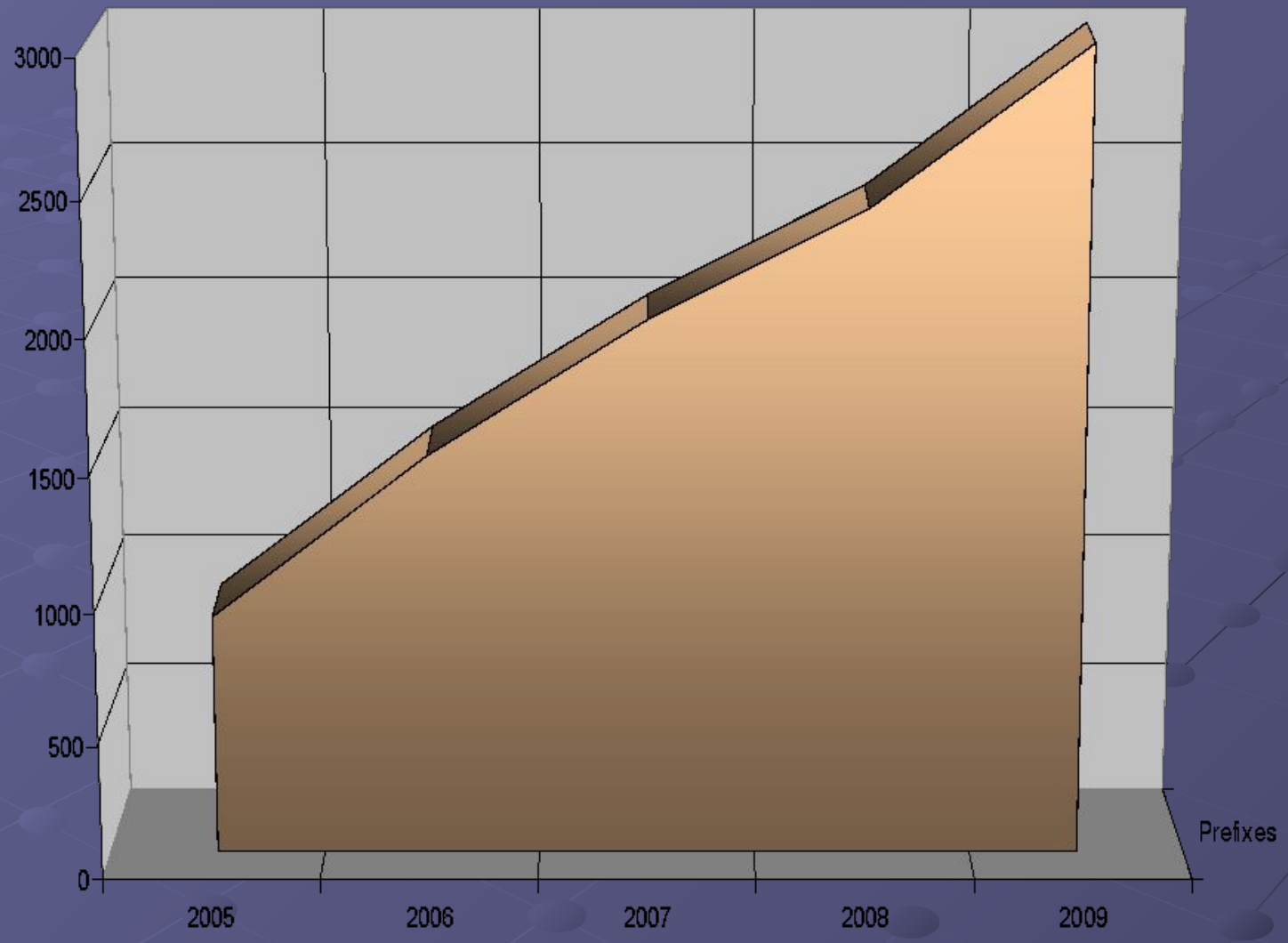
Май 2009

Traffic, Gbps



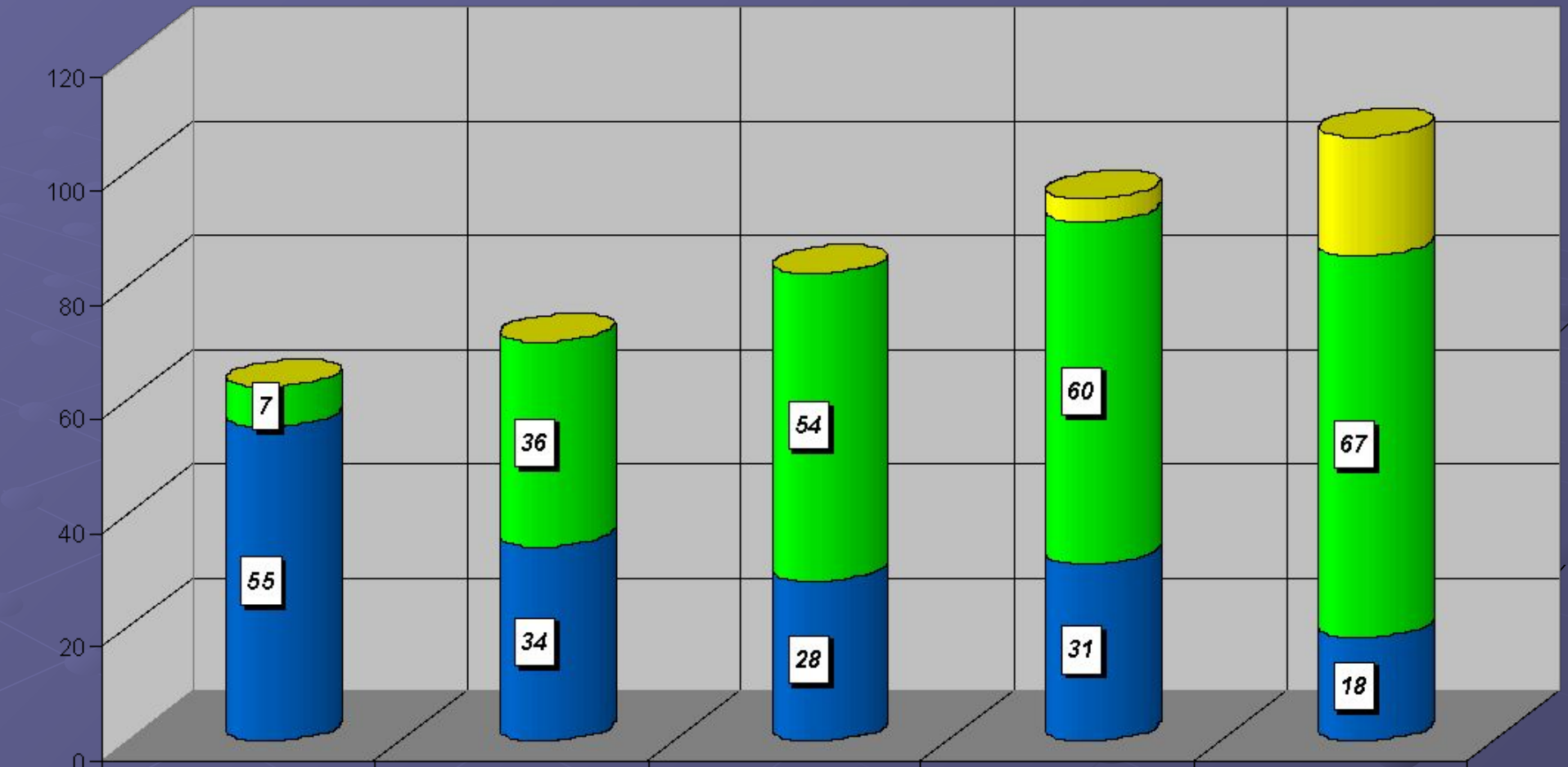
	2005	2006	2007	2008	2009
Traffic, Gbps	1	3	9	15	43

Prefixes



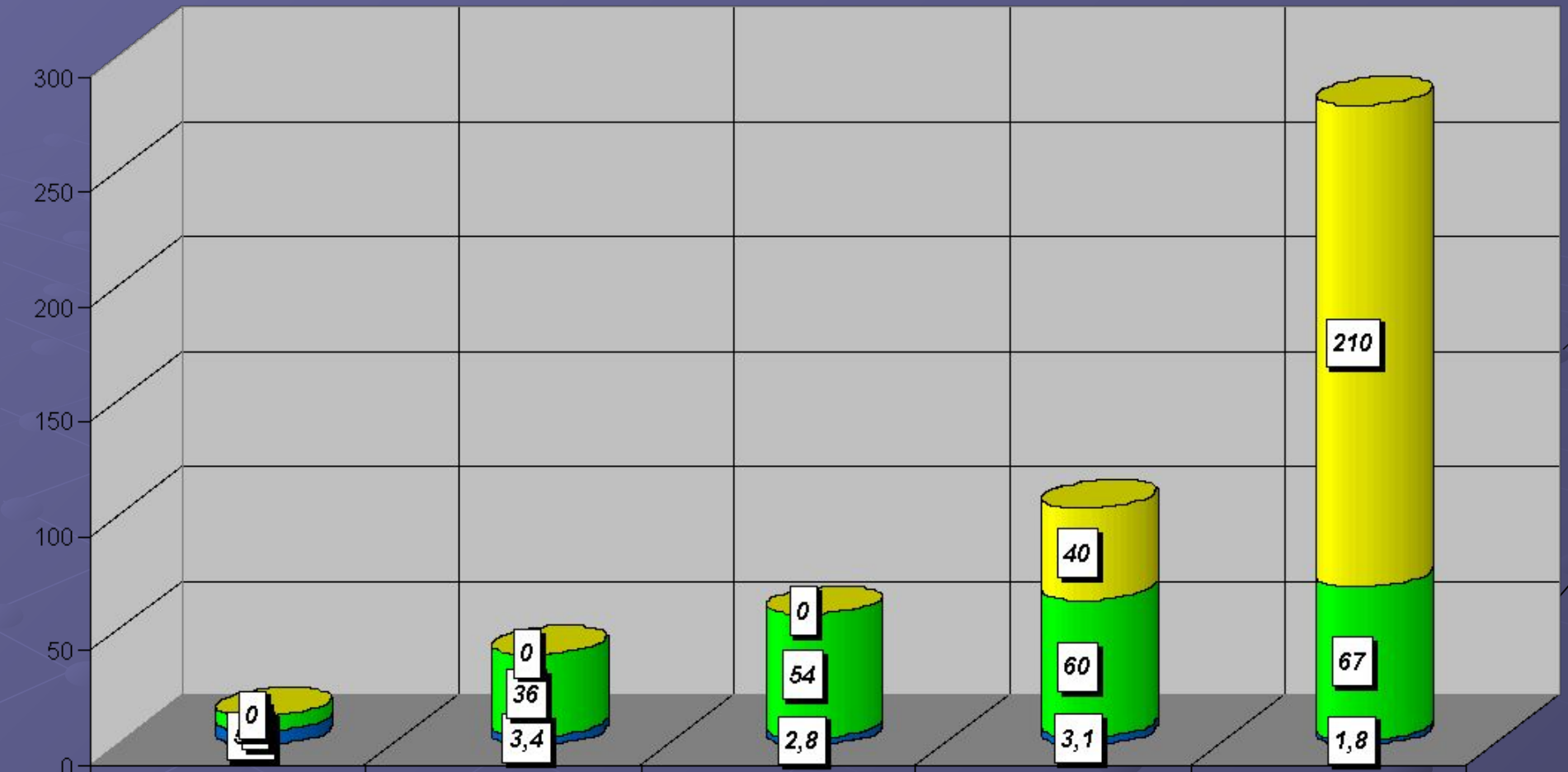
	2005	2006	2007	2008	2009
Prefixes	900	1500	2000	2400	3000

Quantity of UA-IX customer ports



	2005	2006	2007	2008	2009
10GE	0	0	0	4	21
GE	7	36	54	60	67
FE	55	34	28	31	18

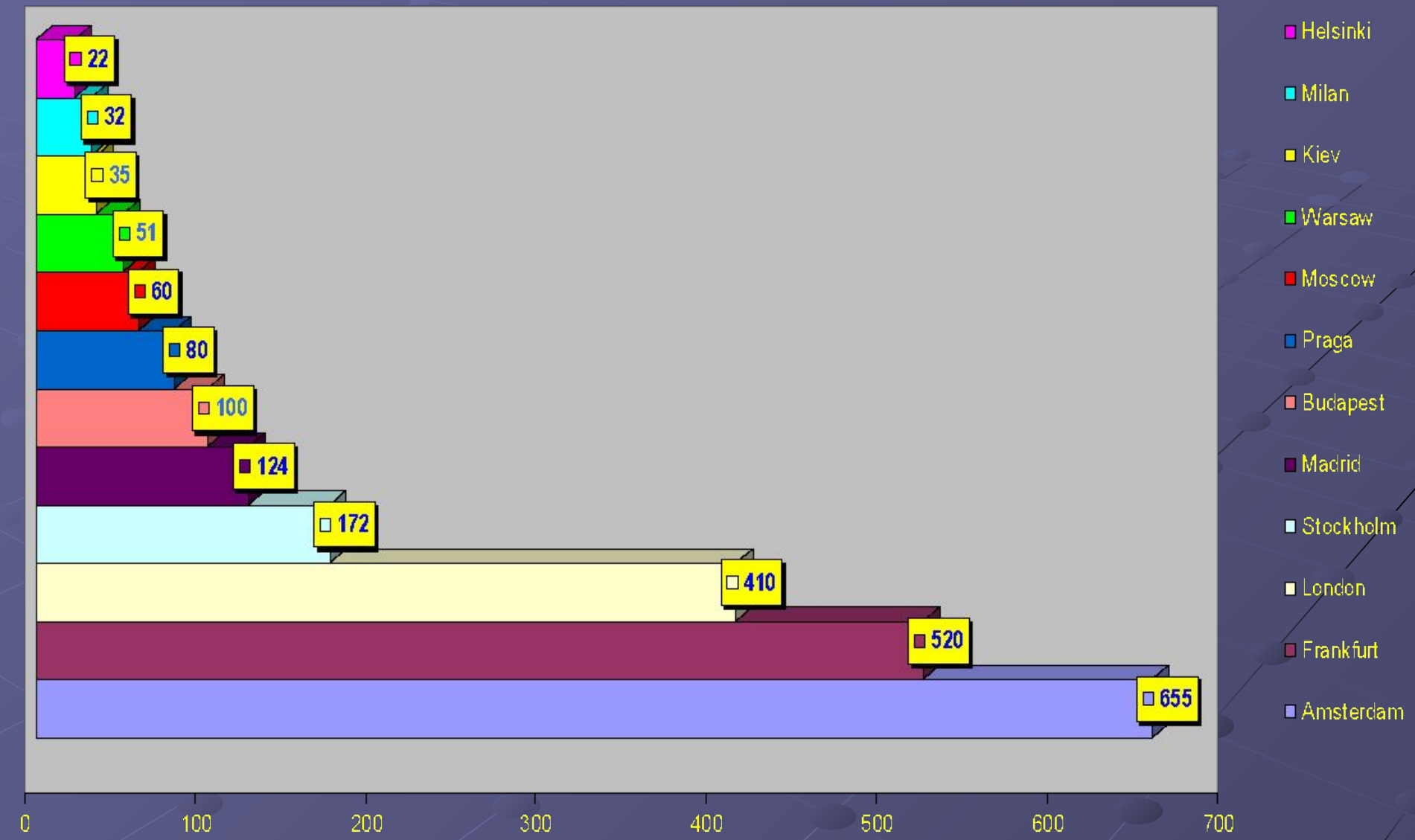
UA-IX customer port bandwidth, Gbps



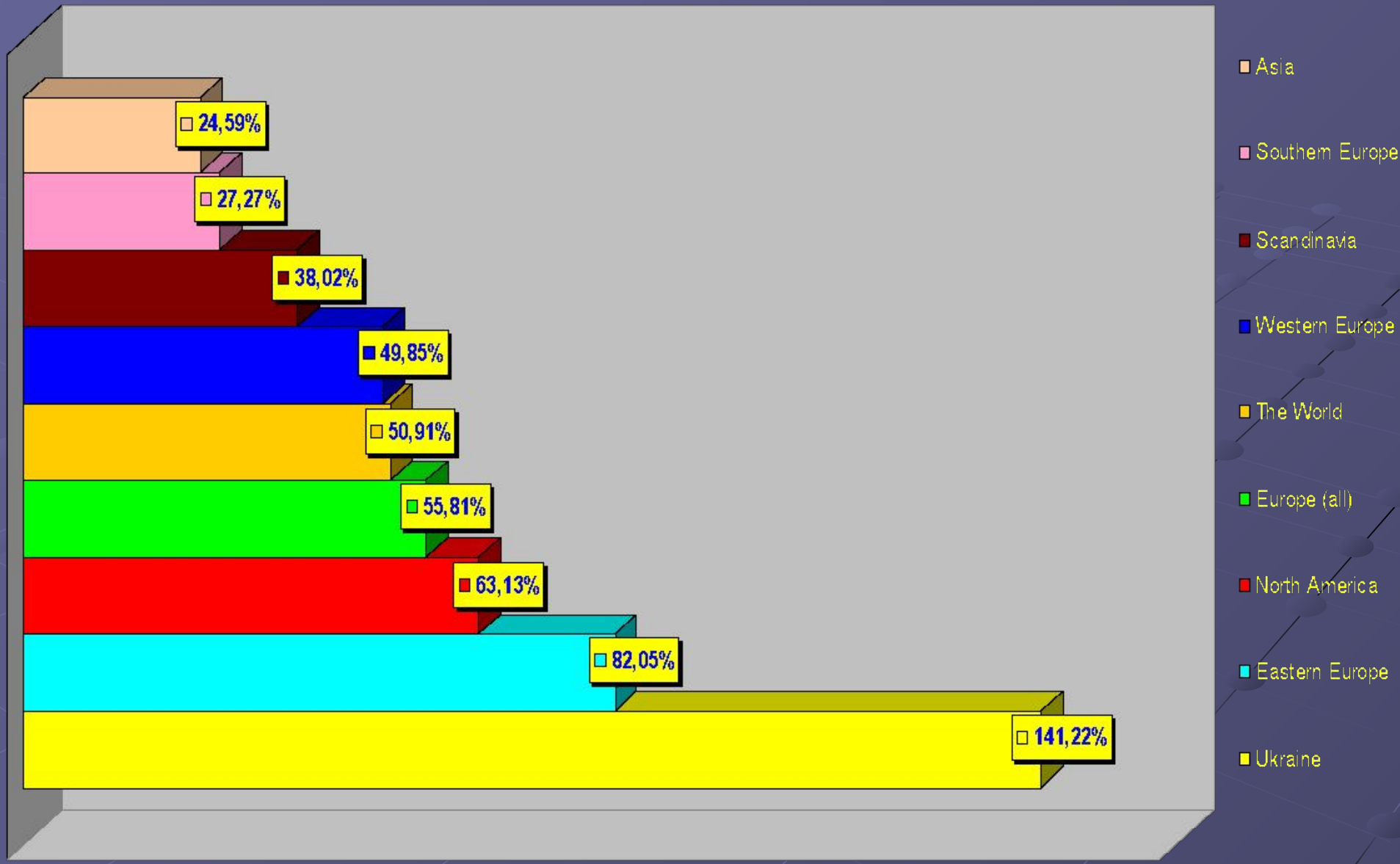
	2005	2006	2007	2008	2009
10GE	0	0	0	40	210
GE	7	36	54	60	67
FE	5,5	3,4	2,8	3,1	1,8

IX и Европа

Пиковая загрузка 11.02.2009, Gbps (105 IXP в 31 стране Европы)



Рост трафика (08.07-08.08)



2009 monthly fees

	UA-IX	PLIX	MSK	LINX	AMS-IX	DE-CIX
100M	€122		€257	€396	€500	€725
1G	€122	€225	€617	€595	€800	€1125
10G	€122	€900	€1851	€1637	€2500	€3425

Цена 10GE в UA-IX (13 копеек за мегабит) дешевле, чем 100M FE
в любой Европейской точке обмена трафиком.
Это одна из причин феномена «бесплатного» украинского трафика.

Аналогичное дорогостоящее оборудование используется в:
LINX - Лондон, CIXP - Женева, FICIX - Хельсинки.

** - перевод валют в Euro выполнен с использованием Bloomberg calculator 10/02/2009.*

Основные результаты за год

- Понижена себестоимость подключения на скорости 10GE путем установки дополнительного коммутатора X650 на 24 порта SFP+. SFP+ дешевле XFP в 1.5 раза, а на расстоянии до 10 м – более, чем в10 раз!
- Повышена надежность сети путем резервирования основных элементов (2xBD-8810, 5xPS, 2xUPS, 2xRS, Nx10GE каналы, для BD-8810 все модули зарезервированы)
- 10GE стал таким же привычным как 1GE. Число включений 10GE превышает число 100FE
- Защита каналов по протоколу EAPS (<50ms recovery)
- Работает эксперимент IPv6
- Интегрирована поддержка 32bit ASN
- Продолжается устойчивый рост трафика в сети

Спасибо за внимание!

Приятного аппетита!