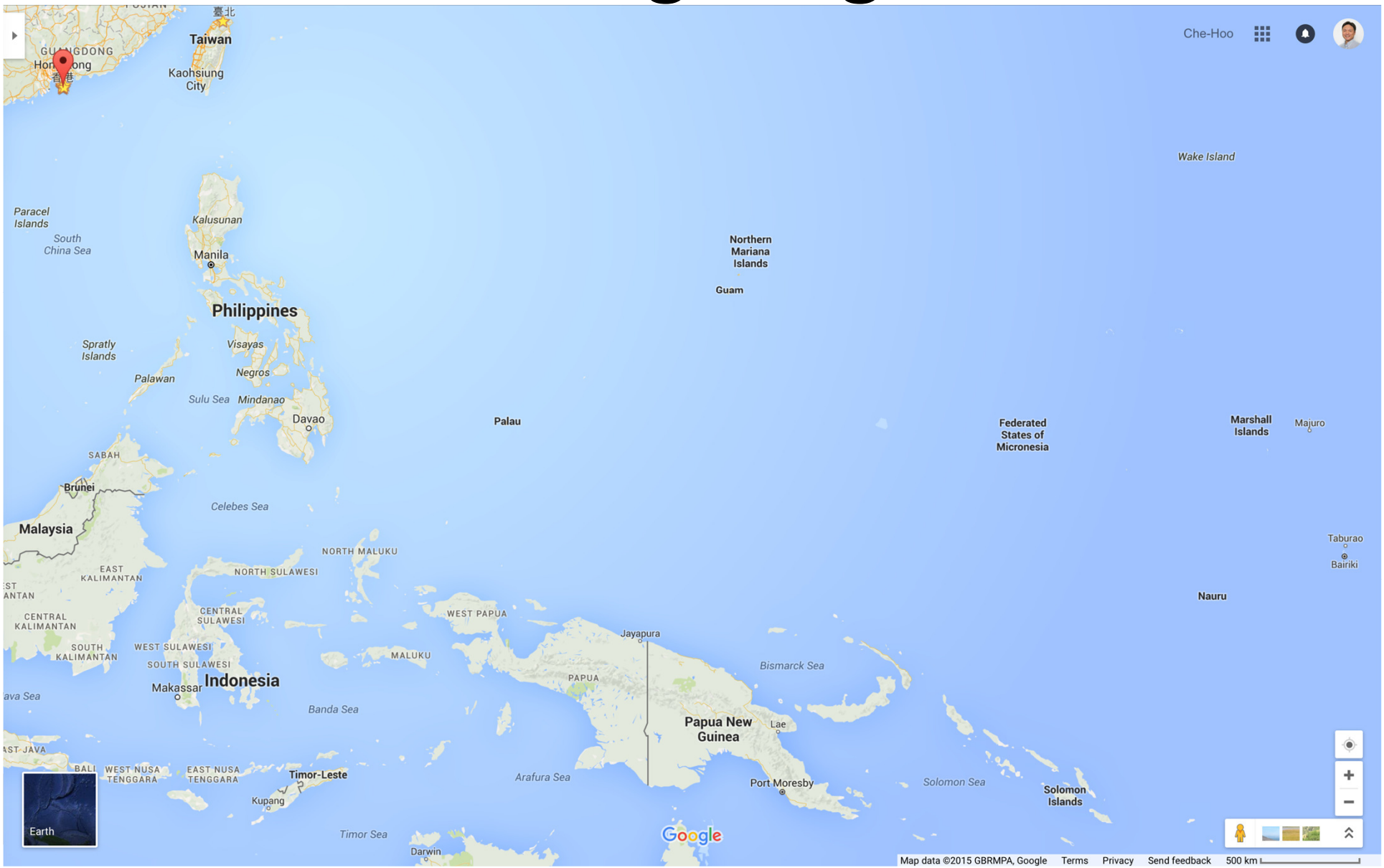
HKIX Sharing at PacNOG 18

Che-Hoo CHENG
CUHK/HKIX
30 Nov 2015

Hong Kong


Che-Hoo
Taiwan
Kaohsiung City
GUANGDONG
Hong Kong
香港
臺北
Wake Island
Paracel Islands
South China Sea
Kalusunan
Manila
Northern Mariana Islands
Guam
Philippines
Spratly Islands
Visayas
Palawan
Negros
Sulu Sea
Mindanao
Davao
Palau
Federated States of Micronesia
Marshall Islands
Majuro
SABAH
Brunei
Malaysia
Celebes Sea
NORTH MALUKU
NORTH SULAWESI
EAST KALIMANTAN
CENTRAL KALIMANTAN
CENTRAL SULAWESI
WEST PAPUA
Jayapura
SOUTH KALIMANTAN
WEST SULAWESI
SOUTH SULAWESI
MALUKU
PAPUA
Bismarck Sea
Makassar
Indonesia
Banda Sea
Papua New Guinea
Lae
Taburao
Bairiki
Nauru
BALI
WEST NUSA TENGGARA
EAST NUSA TENGGARA
Timor-Leste
Kupang
Arafura Sea
Port Moresby
Solomon Sea
Solomon Islands
Timor Sea
Darwin
Earth
Google
Map data ©2015 GBRMPA, Google
Terms
Privacy
Send feedback
500 km

What is HKIX?

- Hong Kong Internet eXchange (HKIX) is a public Internet Exchange Point (IXP) in Hong Kong
- HKIX is the main IXP in HK where various networks can interconnect with one another and exchange traffic
  - Not for connecting to the whole Internet
  - For local peering and regional peering
- HKIX was a project initiated by ITSC (Information Technology Services Centre) of CUHK (The Chinese University of Hong Kong) and supported by CUHK in Apr 1995 as a community service
  - Still fully supported and operated by CUHK
  - 20th Anniversary
- **HKIX serves both commercial networks and R&E networks**
- The original goal is to keep intra-HongKong traffic within Hong Kong

20th Anniversary of HKIX

- HKIX started with thin coaxial cables **in Apr 1995**
  - *Gradually changed to UTP cables / fibers with switch(es)*
    - *low-end -> high-end*
    - *One switch -> multiple switches*
- Participants had to put co-located routers at HKIX sites in order to connect
  - *Until Metro Ethernet became popular*
- It was a free service
  - *Now a fully chargeable service for long-term sustainability*

Requirements to Connect

- Have AS (Autonomous System) number and IP address block(s) allocated/assigned by RIR (Regional Internet Registry)
- Be able to run BGP4 routing protocol
- Have global Internet connectivity independent of HKIX facilities
- Provide its own local circuit(s) to HKIX Access Switch(es)
- Agree to do MLPA for Hong Kong routes

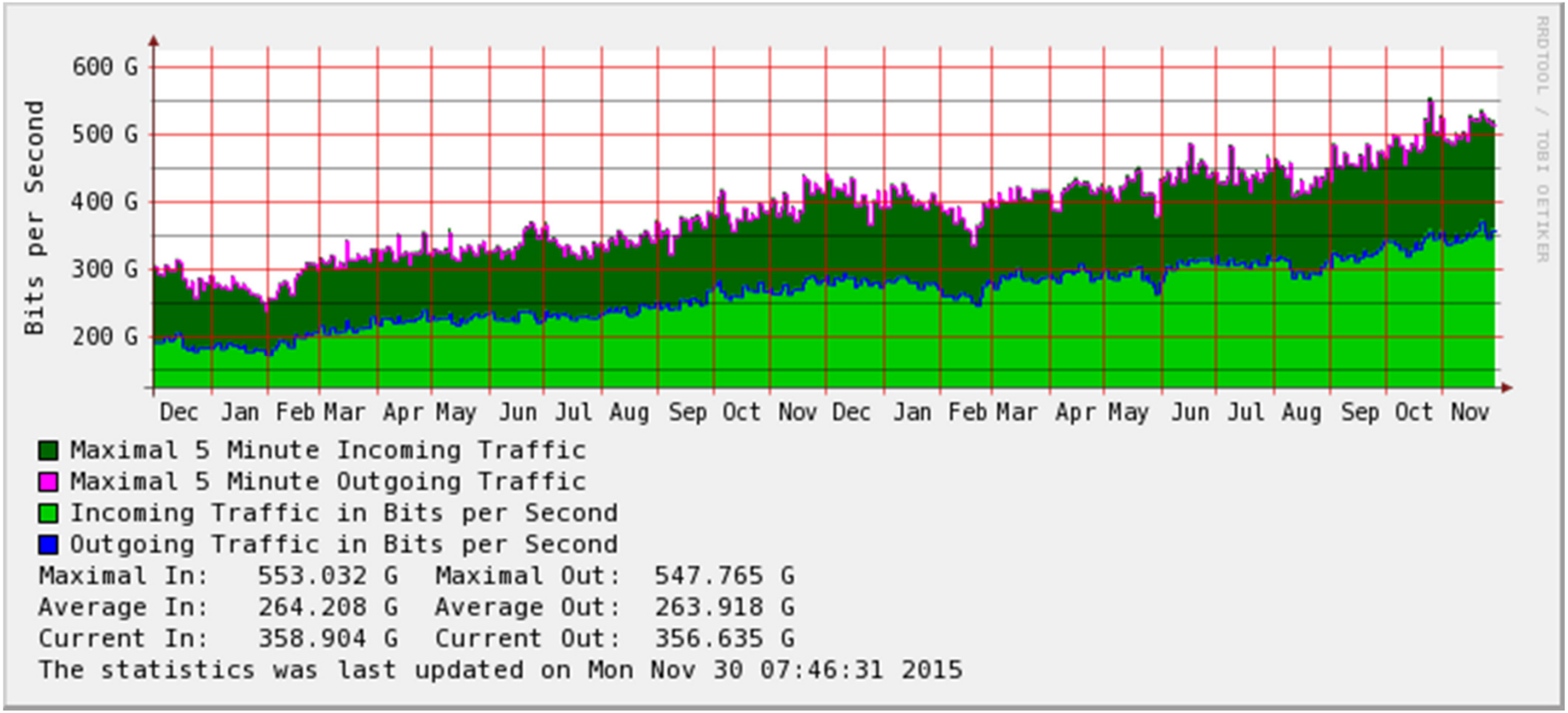
Help Keep Intra-Asia Traffic within Asia

- We have almost all the Hong Kong networks
  - We are confident to say we help keep 98% of intra-Hongkong traffic within Hong Kong
- So, we can attract participants from Mainland China, Taiwan, Korea, Japan, Singapore, Malaysia, Thailand, Indonesia, Philippines, Vietnam, India, Bhutan and other Asian countries
- We now have more non-HK routes than HK routes
  - On our MLPA route servers
  - Even more non-HK routes over BLPA
- We do help keep intra-Asia traffic within Asia
- In terms of network latency, Hong Kong is a good central location in Asia
  - ~50ms to Tokyo
  - ~30ms to Singapore
- So, HKIX is good for intra-Asia traffic
- HKIX does help HK maintain as one of the Internet hubs in Asia

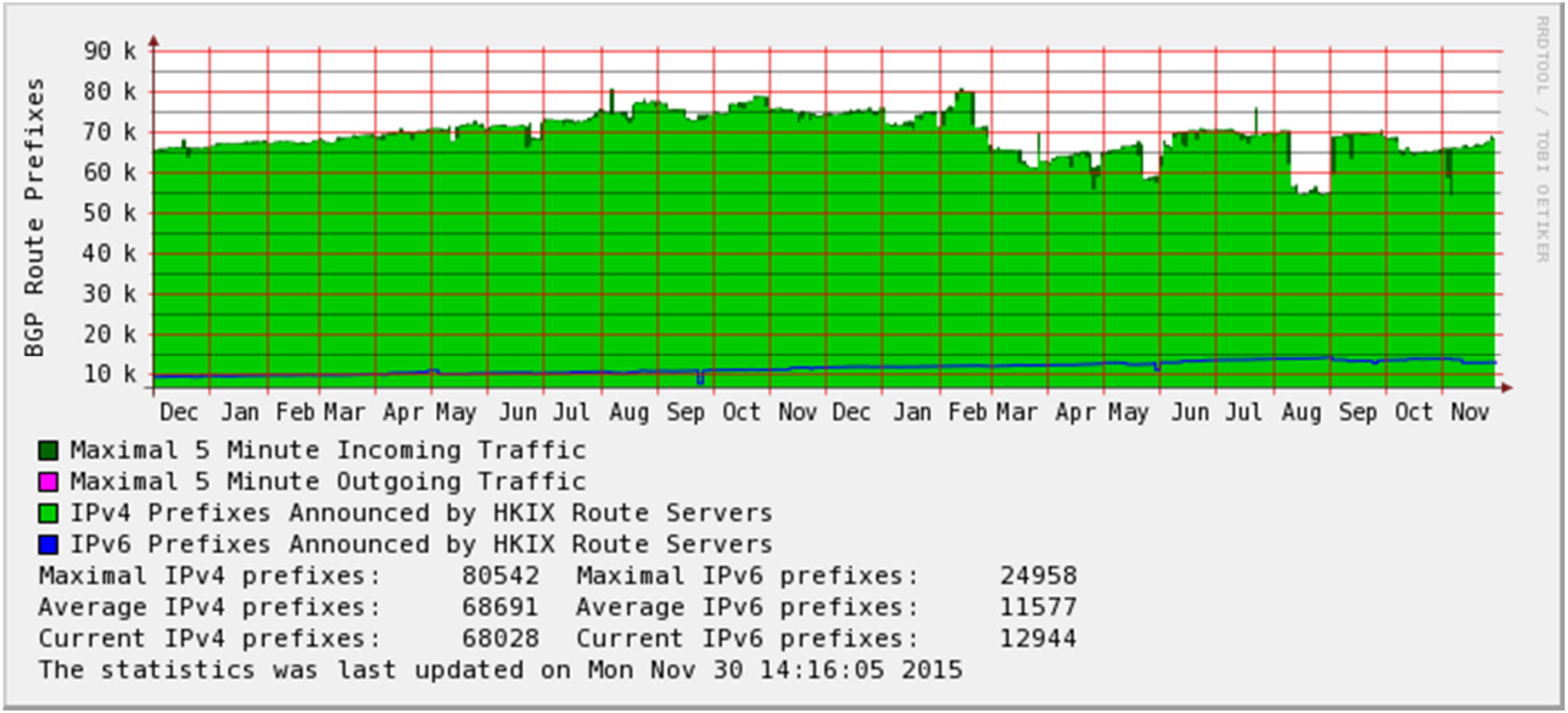
HKIX Today

- Supports both MLPA (Multilateral Peering) and BLPA (Bilateral Peering) over layer 2
- Supports IPv4/IPv6 dual-stack
- Neutral among ISPs / telcos / local loop providers / data centers / content providers / cloud services providers
- More and more non-HK participants
- >230 AS'es connected
- >420 connections in total
  - **>205 x 10GE** + >215 x GE
- ~553Gbps (5-min) total traffic at peak
- Annual Traffic Growth = 30% to 40%

Yearly Traffic Statistics



Number of Routes on MLPA Route Servers (AS4635)



Charging Model

- An **evolution** from free-of-charge model adopted at the very beginning, to penalty-based charging model based on traffic volume for curbing abuse, to now simple port charge model for fairness and sustainability
- **Have started simple port charge model since 01 Jan 2013**
  - See http://www.hkix.net/hkix/Charge/ChargeTable.htm
- Still not for profit
  - HKIX Ltd (100% owned by CUHK) to sign agreement with participants
  - Target for fully self-sustained operations for long-term sustainability

HKIX Charge Table (v1.2)

Standard Port Charges						NRC		MRC	
Port	Interface	Standard Interface	Availability			HKD	*USD*	HKD	*USD*
			HKIX1	HKIX1b	Satellite Sites				
GE	T	Yes	✔			Waived		**936**	*120*
	SX	Yes			✔				
	LX/LH	Yes		✔	✔				
	EX	No		✔		9,360	*1,200*		
	ZX	No		✔		15,600	*2,000*		
10GE	SR	Yes			✔	**10,140**	*1,300*	**7,800**	*1,000*
	LR	Yes	✔	✔	✔	**17,940**	*2,300*		
	ER	No	✔	✔		39,000	*5,000*		
	ZR	No	✔	✔		62,400	*8,000*		
100GE	LR4	Yes	✔	✔	Some	**117,000**	*15,000*	**46,800**	*6,000*
	ER4-Lite	No	✔	✔		468,000	*60,000*		

** Satellite Sites are to be named as HKIX2/3/4/5/6/etc which will be announced soon*
*** E/FE(10ME/100ME) connections are no longer supported*
**** There may be long lead-time for non-standard interfaces (GE-EX, GE-ZX, 10GE-ER, 10GE-ZR and 100GE-ER4-Lite)*
***** The port charges listed do **NOT** cover local circuit/loop charges, cross-connect charges, satellite-site special connection charges, or any other charges needed for making the connection*

Save-IP Discount (applied **ONLY** if IP address is **NOT** needed for the port)		Reduction of MRC for each port entitled	
Port	Conditions	HKD	*USD*
10GE	With Link Aggregation over multiple ports; **NOT** applied to the 1st port which needs IP address	**-780**	*-100*
100GE	With Link Aggregation over multiple ports; **NOT** applied to the 1st port which needs IP address	**-4,680**	*-600*

** No such discount for GE connections and NRC*
*** With Link Aggregation over multiple ports, only card resilience can be provided but not chassis resilience and site resilience*

Volume Discount (applied under the same AS and the same contract **ONLY**)		Reduction of MRC for each port entitled	
Port	Conditions	HKD	*USD*
10GE	Applied to the **5th 10GE port** and onwards	**-780**	*-100*
100GE	Applied to the **3rd 100GE port** and onwards	**-4,680**	*-600*

** No such discount for GE connections and NRC*

REMARKS:
***NRC** = Non-Recurring Charge (**NON**-refundable & **NON**-transferrable to other AS or other company under different name)*
***MRC** = Monthly Recurring Charge*

Updated on 30 Jun 2015

Why HKIX is successful

- Neutral
  - Treat all partners equal, big or small
  - Neutral among ISPs / telcos / local loop providers / data centers / content providers / cloud services providers
- Trustable
  - Fair and consistent
  - Respect business secrets of every partner / participant
- Not for Profit
- *HKIX started very early, well before incumbent telcos started to do ISP business*

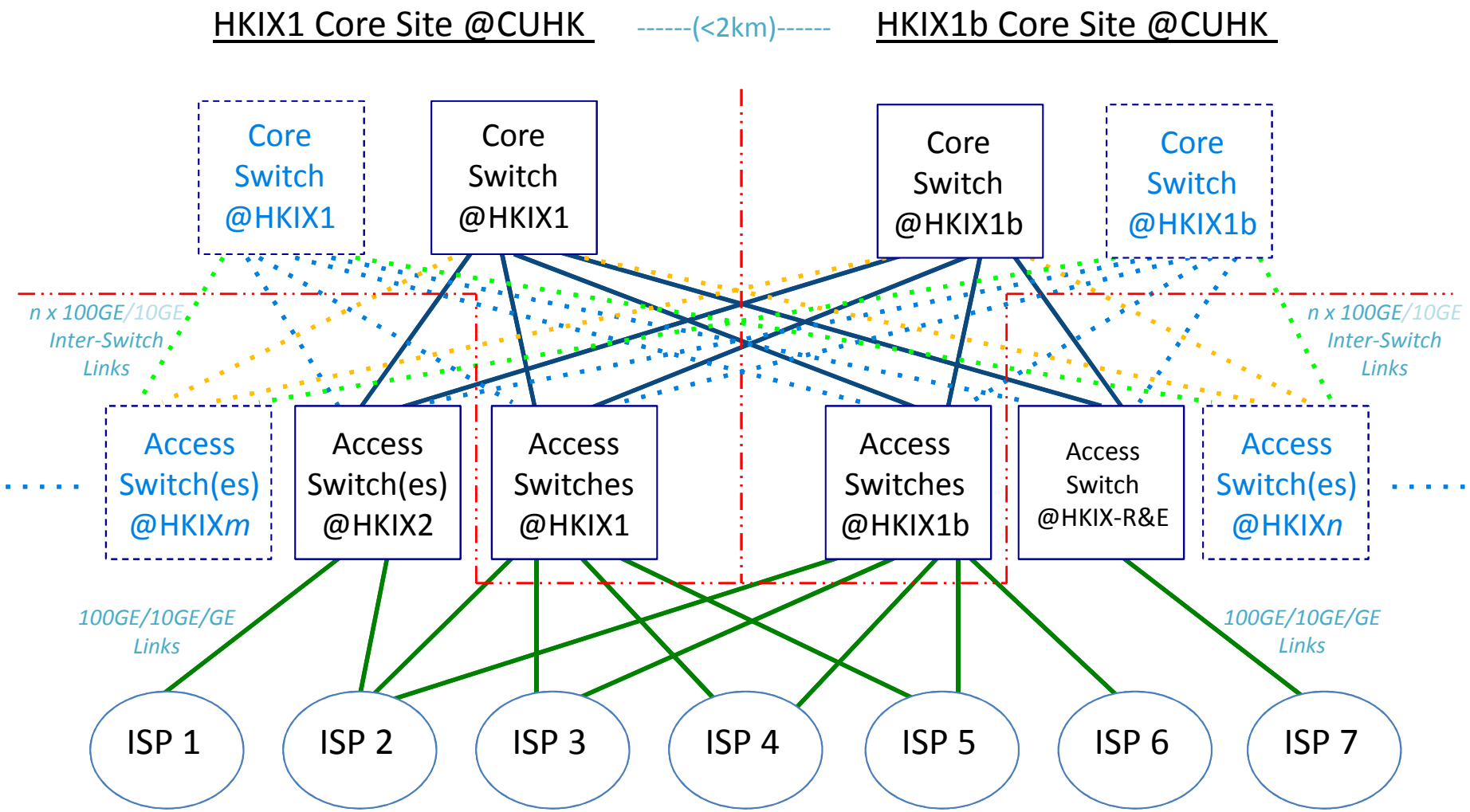
The Upgrade Done in 2014

- Previous single-star topology was not considerable scalable and resilient enough to support future growth of HKIX
- A new highly-scalable two-tier dual-core spine-and-leaf architecture within CUHK was established in 2014 by taking advantage of the new data center inside CUHK Campus
  - HKIX1 site + HKIX1b site as Core Sites
    - Fiber distance between 2 Core Sites: <2km
  - Provide site/chassis/card resilience
  - Support 100GE connections
  - Scalable to support >6.4Tbps total traffic using 100GE backbone links primarily and FabricPath
- **Ready to support HKIX2/3/4/5/6/etc as Satellite Sites**
  - **Satellite Sites have Access Switches only, which connect to Core Switches at both Core Sites**

The Design

- Dual-Core Two-Tier Spine-and-Leaf Design for high scalability
  - Have to sustain the growth in the next 5+ years (to support >6.4Tbps traffic level)
  - Core Switches at 2 Core Sites (HKIX1 & HKIX1b) only
    - No interconnections among core switches
  - Access Switches to serve connections from participants at HKIX1 & HKIX1b
    - Also at Satellite Sites HKIX2/3/4/5/6/etc
    - Little over-subscription between each access switch and the core switches
  - FabricPath (TRILL-like) used among the switches for resilience and load balancing
- Card/Chassis/Site Resilience
  - LACP not supported across chassis though (card resilience only)
- 100GE optics support
  - LR4 for <=10km and ER4-lite for <=25km (1Q2016)
  - Support by local loop providers is key
- Port Security still maintained (over LACP too)
  - Only allows one MAC address / one IPv4 address / one IPv6 address per port (physical or virtual)
- Have better control of Unknown-Unicast-Flooding traffic and other storm control

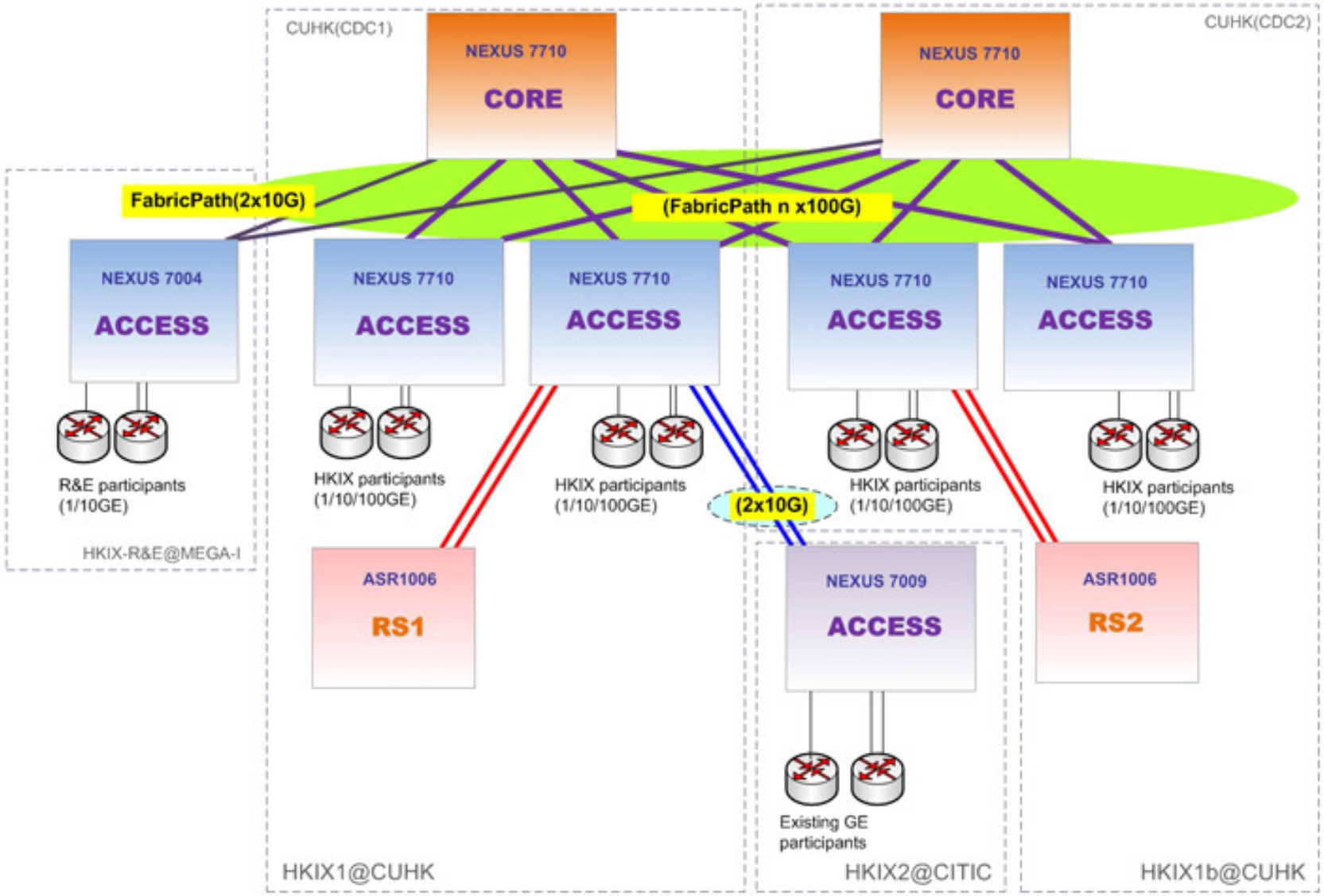
New HKIX Dual-Core Two-Tier Spine-and-Leaf Architecture For 2014 and Beyond



The Migration

- New switches in production at HKIX1 site starting Mar 2014
  - While HKIX1b site was still under construction
  - Interconnected with the old core switch with n x 100Gbps (n=2 and then 4) during the migration period
  - Parellel run with old switches for supporting connections
- All new connections on new access switches since Mar 2014
  - While existing connections were being moved to the new access switches one by one
- By early Aug 2014, all 10GE connections had been moved to new access switches
- HKIX1b site was put into production in Nov 2014 to provide true site resilience
- All remaining GE connections had also been moved to new access switches
  - Deadline was 30 Jun 2015
  - No E/FE support starting then
  - The remaining old switch had been decommissioned
- RS1/RS2 & HKIX-R&E have been moved to the new architecture
- HKIX2 will be moved to the new architecture as Satellite Site in due course

HKIX Network Diagram (JUL 2015)



1. HKIX1 and HKIX1b are the two core sites of HKIX. HKIX participants are encouraged to connect to both sites for site resilience.
2. For those participants which connect to only one of the two core sites, HKIX1b instead of HKIX1 will be assigned for better load balancing.

One Very Critical Point for an IXP

- An IXP must NOT be vulnerable to DDoS attack itself
- Congestion at one port must NOT cause trouble to any other ports

FabricPath

Being Used in New Architecture

- We adopt spine-and-leaf architecture for high scalability
  - Avoid connecting participant ports on core switches
- The Spanning Tree Protocol (STP) domains do not cross into the FabricPath network
  - Layer 2 gateway switches, which are on the edge between the CE and the FabricPath network, must be the root for all STP domains that are connected to a FabricPath network
- Load balancing is working fine
  - Even with odd number of links
- Transparent to participants (i.e. no BGP down) when adding/removing inter-switch links

IPv4 Address Renumbering and Route Servers Upgrade

Migration Date: 12-15 Jun 2015 (Fri-Mon)

IPv4 Address Renumbering

- Network mask was changed to /21 from /23, for accommodating future growth
- ALL participants had to change to **NEW 123.255.88/21**, away from *OLD 202.40.160/23*
- **Parallel run of old and new IPv4 addresses only during the 4-day migration period, having learnt from experience of other IXPs**
- MLPA: New route servers support new IPv4 addresses while old route servers supported old addresses, but IPv6 was handled separately
- BLPA: Individual participants had to coordinate with their peering partners directly
- *No change to IPv6 addresses*

Route Servers Upgrade

- The two old route servers were decommissioned at the end of the period
- Two new route servers had been installed at HKIX1 and HKIX1b (the two HKIX core sites)
- More route server features will be supported later

IPv4 Address Renumbering and Route Servers Upgrade

Considerations beforehand:

- Peak traffic level: ~450Gbps (5-min average)
- # of prefixes on MLPA route servers: ~80K IPv4 prefixes & ~12K IPv6 prefixes
- Complexity of migration: participants from many different time zones & 330+ BGP sessions
- Have to minimize topology changes and configuration changes to participants
- Also need to care about bilateral peering – both old and new networks on the same VLAN
- Have to take care of capacity requirements and routing performance if transit is to be provided between old and new networks

Three options had been looked into:

- **Big Bang Approach** – Pros: Minimum effort to HKIX / Cons: Need coordination with ALL participants for aligning the maintenance window which is extremely difficult
- **Parallel Run with Transit** – Pros: Easier for participants / Cons: Transit routers would need to handle huge traffic of up to 300Gbps and would not be able to support BLPA across old and new networks
- **Parallel Run with Secondary Address** – Pros: Flexible changing time as secondary address can be configured before migration / Cons: Participants need to configure 2nd address on all the router interfaces connecting to HKIX

IPv4 Address Renumbering and Route Servers Upgrade

After careful studies and making reference to other IXPs around the world, we finally decided to take the approach of **Parallel Run with Secondary Address** + **Transit Router** (for backup and contingency) and do the renumbering within **4-day period** (Fri to Mon)

IPv4 Address Renumbering and Route Servers Upgrade

Communication Part:

Before Migration

- 3-4 months – Announced the address renumbering at APRICOT-APAN 2015 and then HKNOG 1.1
- 3 months – Made the announcement by emails to all HKIX participants without detailed info and requested them to provide their contact points for the IPv4 renumbering tasks
- 2-3 months – Replied acknowledged participants and let them know that a migration webpage had been established and the latest information would be published there
- 2 months – Sent reminders to participants who had not respond through all contact points (i.e. contractual / billing / technical contacts) as their commitment to the address renumbering would be very important to the whole project
- 5 weeks – Provided the information of new IP addresses and published the mapping of old address to new address on the migration webpage
- 4 weeks – Published final schedule, migration details, sample configurations and FAQs
- 3 weeks – Sent individual emails to participants and asked them to confirm and specify the intended migration time within the 4-day period
- 1-2 weeks – Followed up again if reply was still not received from the participants
- 1 week – Set up the Command and Control Center (CCC) and ensure that all email templates were ready in place
- 1 day – CCC in operations, 24-hour technical team standby

IPv4 Address Renumbering and Route Servers Upgrade

Communication Part:

During Migration

- Closely monitored the migration progress and escalated the cases to technical team in case problem reported by HKIX participants
- Provided the latest renumbering status on migration webpage and let participants know the up-to-date progress

After Migration

- Followed up with participants and requested them to remove the old addresses from their router interfaces

IPv4 Address Renumbering and Route Servers Upgrade

Technical Part:

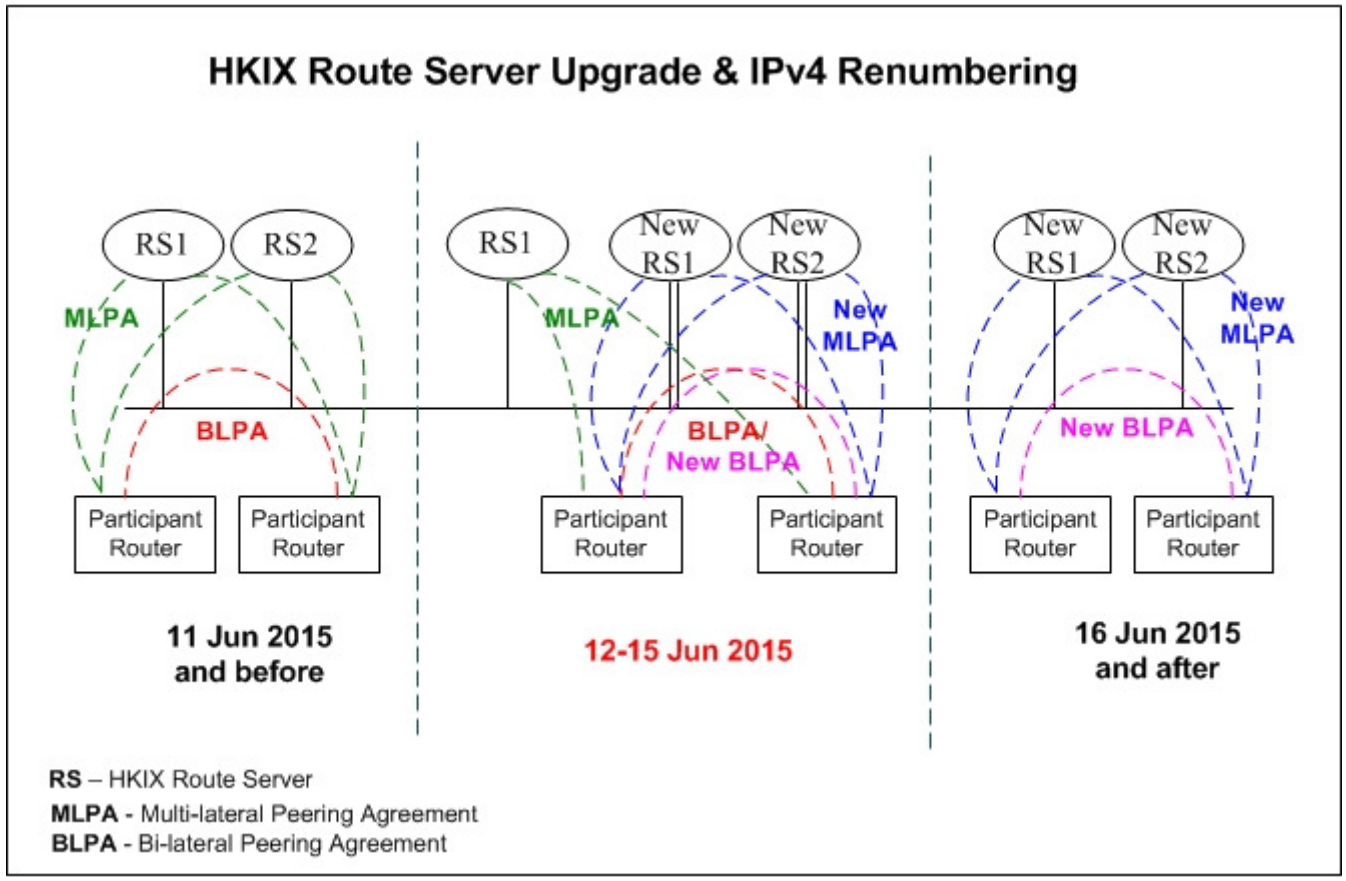
Before Migration

- 2 months – Tested the equipment in lab and did simulation with different scenarios
- 1 month – Equipment trial run and final acceptance test
- 3 weeks – Installed the new route servers & backup transit router
- 2 weeks – Replaced RS2 with new route server using OLD address
- 1 week – Deployed new RS1 and invited some participants for pilot testing

During Migration

- Start of Day 1 – Re-configured RS2 to use NEW address; Set up new RS1 with NEW address; old RS1 still in production
- Day 1-4 – Set up BGP sessions with participants on new RS1 & RS2; Parallel run of new and old Route Servers
- Day 1-4 – Monitored the traffic and the overall progress with the migration schedule provided by participants
- Day 1-4 – Provided instant technical assistance (including trouble-shooting) to participants in case they had difficulties in setting up the BGP sessions
- Day 1-4 – No observable traffic drop during the period
- End of Day 4 – Shut down and decommissioned old RS1 & RS2

IPv4 Address Renumbering and Route Servers Upgrade



IPv4 Address Renumbering and Route Servers Upgrade

Lessons Learnt:

- The Key to Success is Good Planning and Good Communication
- Parallel run for migration is a must but there is no need to do parallel run for too long
- Making contact with all the participants is most time-consuming but is also most important

Many thanks to the whole HKIX Team and all the HKIX participants involved

Setting up
Multiple HKIX Satellite Sites

- Allow participants to **connect to HKIX more easily at lower cost** from those satellite sites in Hong Kong
- Open to all commercial data centres in HK which fulfil minimum requirements so as to maintain neutrality which is the key success factor of HKIX
  - ISO27001 requirement
  - Minimum size requirements
  - Requirements on circuits connecting back to the two HKIX core sites
  - Non-exclusive
- Intend to create win-win situation with satellite site collaborators
- To be named HKIX2/3/4/5/6/etc
- *NOTE: HKIX1 and HKIX1b (the two HKIX core sites) will continue to serve participants directly*

Planned Work in 2015-2016

- **Introduce advanced Route Server functions**
- Better Control of Proxy ARP
- Better support for DDoS Mitigation
- More L2 ACL on HKIX peering LAN
- Portal for HKIX participants
  - Port info and traffic statistics
  - Self-service port security update
  - Network maintenance schedule
- Improve after-hour support
- ISO27001

Thank you!