HKIX Upgrade to 100Gbps-Based Two-Tier Architecture — Experience Sharing and Support to R&E Networks

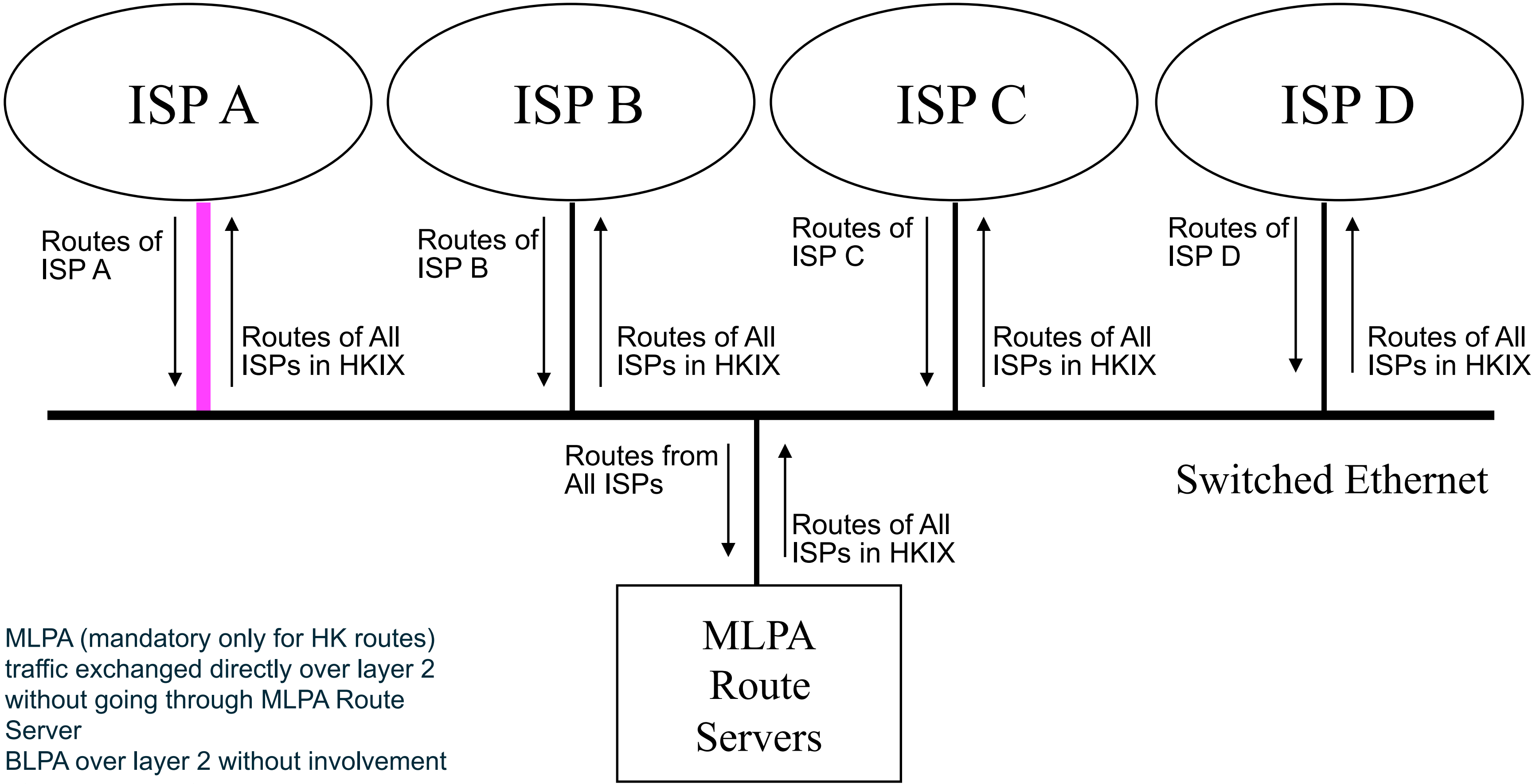
Che-Hoo Cheng
CUHK/HKIX

2014.08.14

What is HKIX?

- 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
- 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
- HKIX serves both commercial networks and **R&E networks**
- The original goal is to keep intra-HongKong traffic within Hong Kong

HKIX Model —
MLPA over Layer 2 + BLPA



- MLPA (mandatory only for HK routes) traffic exchanged directly over layer 2 without going through MLPA Route Server
- BLPA over layer 2 without involvement of MLPA Route Server
- Supports both IPv4 and IPv6 over the same layer 2 infrastructure

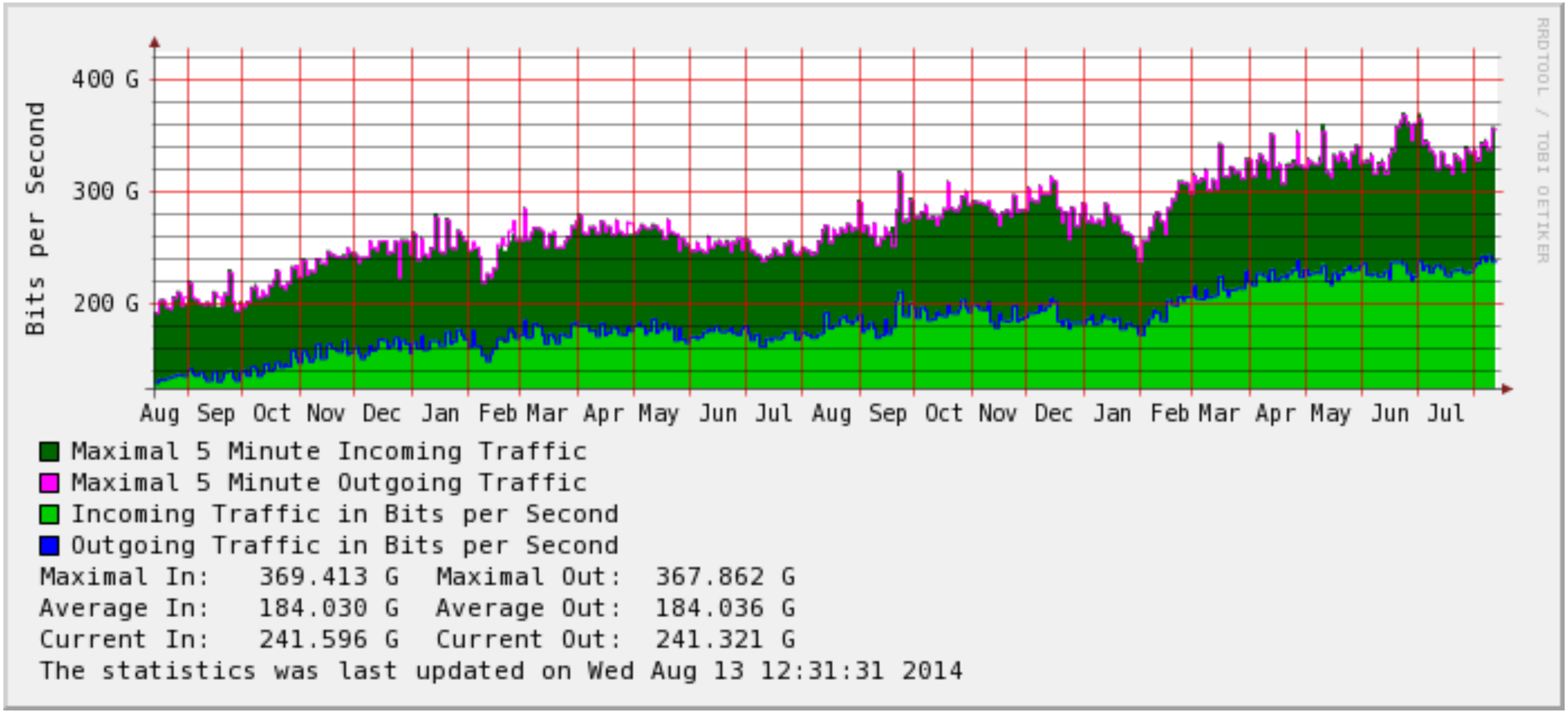
Help Keep Intra-Asia Traffic within Asia

- 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, Qatar 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
- HKIX is good for intra-Asia traffic

HKIX Today

- Supports both MLPA (Multilateral Peering) and BLPA (Bilateral Peering) over layer 2
- Supports IPv4/IPv6 dual-stack
- Accessible by most local loop providers
- Neutral among ISPs / telcos / local loop providers / data centers / content providers / cloud services providers
- More and more non-HK participants
- >240 ASNs connected
- >370 connections in total
  - >130 10GE connections
- ~370Gbps (5-min) total traffic at peak
- Annual Traffic Growth = 30% to 40%

Yearly Traffic Statistics



🟩 Maximal 5 Minute Incoming Traffic
🟪 Maximal 5 Minute Outgoing Traffic
🟩 Incoming Traffic in Bits per Second
🟦 Outgoing Traffic in Bits per Second

```
Maximal In:    369.413 G  Maximal Out:  367.862 G
Average In:    184.030 G  Average Out:  184.036 G
Current In:    241.596 G  Current Out:  241.321 G
The statistics was last updated on Wed Aug 13 12:31:31 2014
```

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**
  - E/FE/GE – US$120/port/month (with no one-time charge)
  - 10GE – US$1,000/port/month (plus one-time charge)
  - See http://www.hkix.net/hkix/Charge/ChargeTable.htm
- Co-location service for strategic partners only is chargeable
- 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

Values of HKIX to Hong Kong

- A key information infrastructure bringing faster and cheaper connectivity to Hong Kong citizens
- A key component for developing Hong Kong as an Internet hub in Asia
- A key component for helping Hong Kong's competitiveness in the cyber world
- A key component in facilitating competition in the telecommunication sector
- Considered as Critical Internet Infrastructure in Hong Kong

HKIX's Advantages

- Neutral
  - – Treat all partners equal, big or small
  - – Accessible by all local loop providers
  - – Neutral among ISPs / telcos / local loop providers / data centers / content providers / cloud services providers
- Trustable
  - – Respect business secrets of every partner / participant
- Not for Profit

2013 and Beyond?

- A lot of new data centers will be in operations in Hong Kong starting 2013
- More and more cloud / content services providers setting up presence in Hong Kong
- What will happen to the industry and the market?

- **HKIX must be well-prepared for the possibly higher growth**

In Need of Continuous Upgrades for HKIX

- Peak total traffic is growing continuously
- Did not have enough ports at HKIX1 for new connections at times
- Need to support 100GE interfaces
- Resilience is becoming a bigger concern to HKIX participants
- **We cannot afford any performance bottleneck**
- **We must cope with the continuous technology changes**

CUHK's Vision

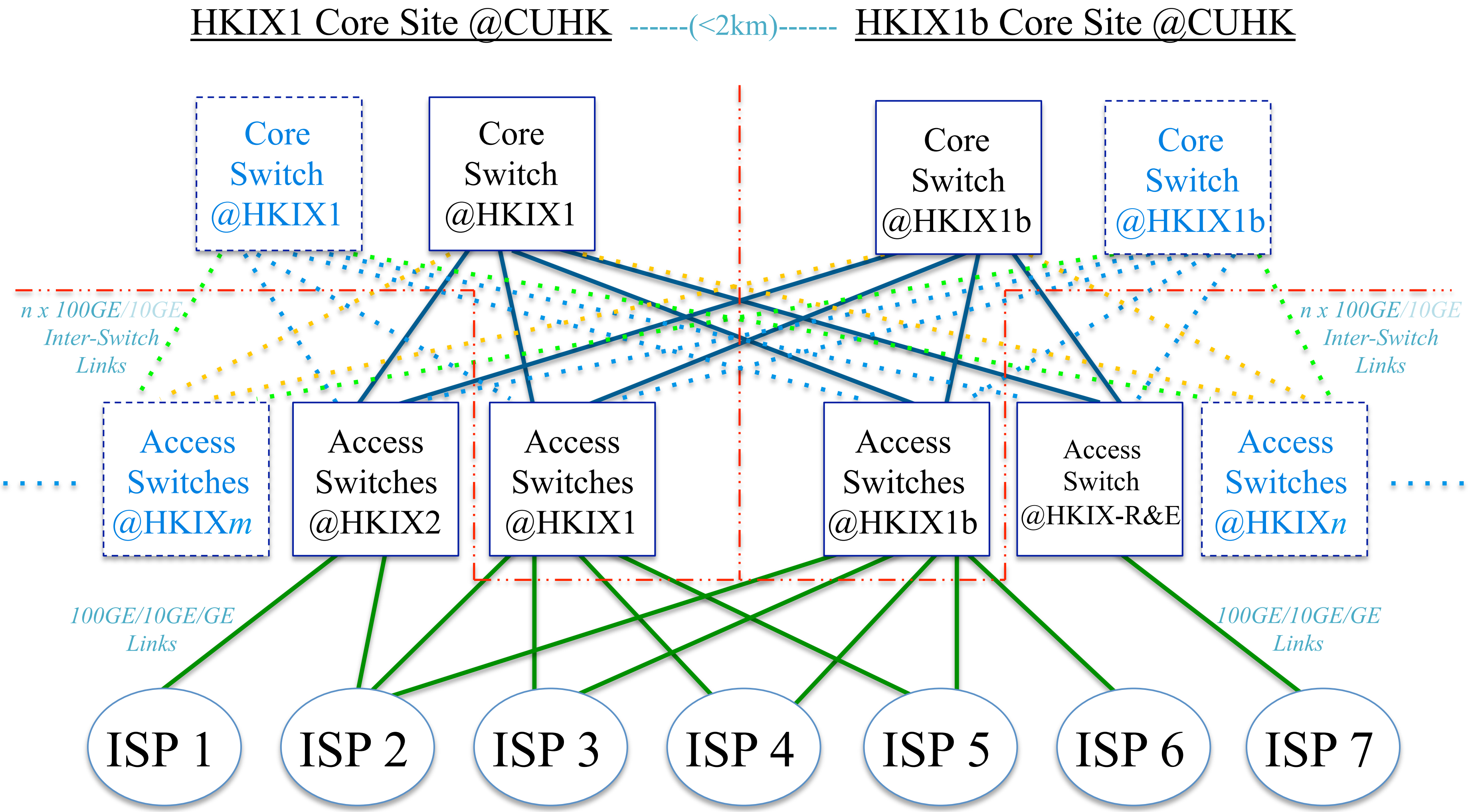
- CUHK has a strategic uniqueness in running HKIX in a long-term
- While CUHK does not have a service provider role, we are still obligated to continue managing it as a public service
- HKIX is very much like road infrastructure and airport in Hong Kong
- Support from HKSAR Government is needed to make it prosper, and to maintain it as an Asian internet hub

- **HKSAR Government has provided one-off funding for capital expenses of network equipment upgrade in 2013-14**

The Plan

- Have started simple port charge model since Jan 2013
  - Maintain as not-for-profit operations
  - Target for fully self-sustained operations for long-term sustainability
- Deploying new highly-scalable two-tier dual-core architecture within CUHK by 2014 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 having Access Switches only which connect to Core Switches at both Core Sites**

HKIX Dual-Core Two-Tier Architecture For 2014 and Beyond



The Design

- Dual-Core Two-Tier 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 for <=40km (1Q2015)
  - 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

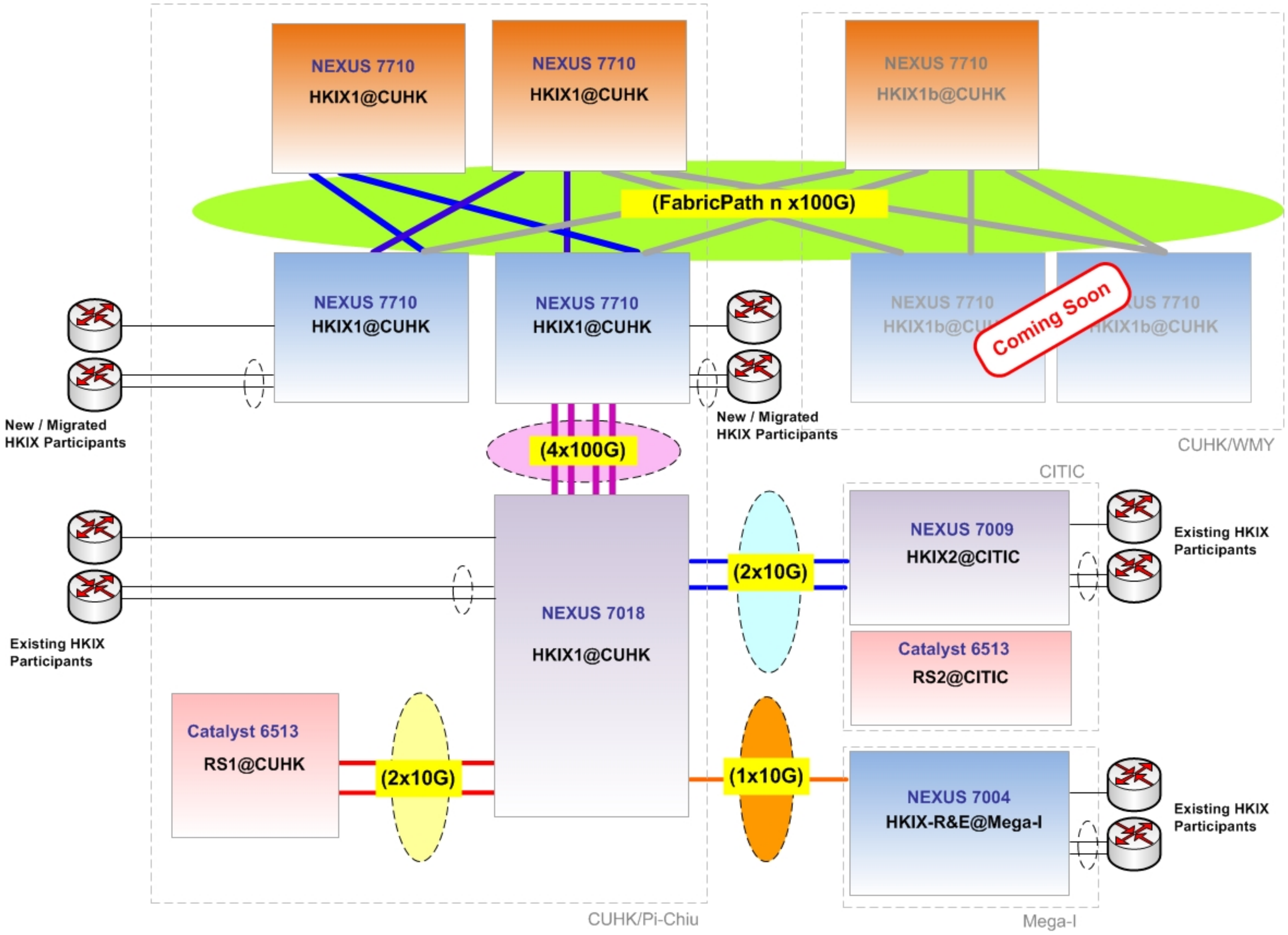
HKIX1b Site Delayed

- Raised Floor System issue
  - Hopefully it will be ready by Oct 2014
- 2 Core Switches at HKIX1 site to start migration first
  - May need to have more access switches at HKIX1
- Population of HKIX1b site will take much longer time so the strategy is:
  - All new connections to be set up at HKIX1b first unless for resilience purpose
  - Half of the existing connections at HKIX1 will be "forced" to moved to HKIX1b when their local loop contracts expire

The Migration

- New switches in production at HKIX1 site starting Mar 2014
  - While HKIX1b site is still under construction
  - Interconnected with the old core 7018 switch with n x 100Gbps (n=2 and then 4) during the migration period
- All new connections are now on new access switches
  - While existing connections are being moved to the new access switches one by one
- **By early Aug 2014, all 10GE connections had been moved**
- Remaining GE connections will be moved gradually
  - Deadline is 30 Jun 2015
  - No E/FE support starting then
- RS1, RS2, HKIX2 & HKIX-R&E will also be moved away from the old architecture to the new architecture soon

HKIX Network Diagram (July 2014)



DDoS Attack During Migration

- Old equipment limitation
  - hashing by source and destination MAC addresses
  - Very high traffic from old to new targeting one single destination MAC address
  - Feedback mechanism to drop packets at sources
- Workaround
  - Layer 2 Netflow (v9) to check the high-volume sources
  - Migrate them to new switches immediately

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
- Network QoS Policy - Congestion Control Mechanisms
  - Default is "Burst optimized" which is not good for IXP because of sharing of buffer by multiple ports
  - "Mesh optimized" is more suitable for IXP

100GE Interfaces

- CPAK instead of CFP
  - 12 ports per line card so can support high density 100GE (line-rate)
    - CFP – only 2 ports per line card
  - SR10
    - MMF/OM3 – up to 100 meters
    - MMF/OM4 – up to150 meters
    - Fibers (24-core MPO cables)
      - Using cheaper cables
        - » ~US$220 for 5-meter & ~US$280 for 10-meter
      - Long delivery lead time
  - LR4
    - SMF – up to 10km
  - ER4
    - SMF – up to 40km
    - Seems more needed than LR4
    - Not available yet, need to wait until 1Q2015
  - Power consumption lower
    - Not hot so greener

10GE SFP+ Transceivers

- Same type of LR transceivers can have Tx Power (optical) difference of up to 2dbm
  - Seems different batches have different Tx Power
  - Record down Tx Power every time for comparison
  - Seems to have down trend
- ZR/ER are also supported mainly for local loops carriers
- LACP mixed with ER & ZR
  - Running ok

Proxy ARP Threat

- Can use Dynamic ARP Inspection (DAI) to maintain static ARP list
  - But not used yet as it is manual
  - Need to input a few commands for this instead of just one command

FabricPath

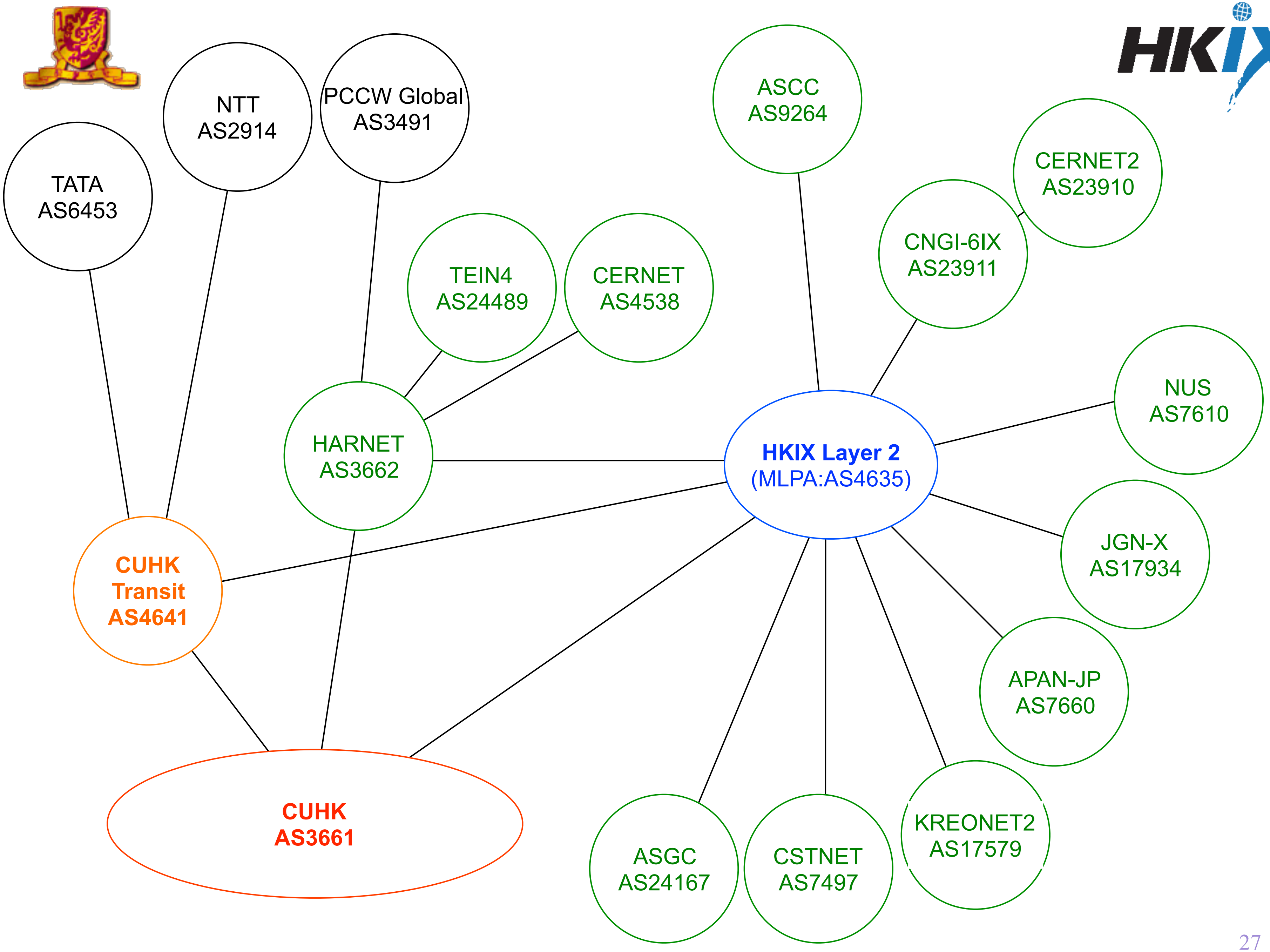
- ISIS neighboring timeout took a few minutes to recover
  - BGP failed
  - Physical issue?
- Load Balancing seems working fine
  - Even with 3 links

Other Problems Seen

- 100GE card in core switches self-reload a few times
- Supervisor Engine (SUP) switch-over not working
  - 'mac packet-classify' on port-channel member interfaces caused ACL manager crashed
  - In case of SUP switchover, will go into a boot loop
  - Workaround is to remove the config at member ports
  - Known bug
  - Same would happen on 7000
    - We were lucky that we did not encounter problems
- LACP cannot mix SR10 and LR4 for 100GE
- Waiting for 6.2.10 available in late Aug which should solve most problems

Other Practices

- Always keep spare chassis/line cards/transceivers on-site for back-up
- FabricPath must use F cards
  - Not to mix M cards and F cards in the same chassis
  - We use only F cards on 7710
  - We still use M cards on 7018 (no FabricPath support)
    - 7004 at HKIX-R&E also uses only F cards so can support FabricPath
- Not to mix F2e-GE/10GE and F3-GE/10GE cards in the same chassis to avoid possible problems
  - No LACP across two different types of cards


TATA
AS6453
NTT
AS2914
PCCW Global
AS3491
ASCC
AS9264
CERNET2
AS23910
CNGI-6IX
AS23911
TEIN4
AS24489
CERNET
AS4538
NUS
AS7610
HARNET
AS3662
HKIX Layer 2
(MLPA:AS4635)
JGN-X
AS17934
CUHK
Transit
AS4641
APAN-JP
AS7660
KREONET2
AS17579
CUHK
AS3661
ASGC
AS24167
CSTNET
AS7497

Special Services for R&E Networks

- Support LACP and Trunk Ports at HKIX-R&E POP
- Jumbo Frame support
- Special VLANs
  - For private interconnections among any 2 R&E networks
  - One special R&E IX-VLAN for interconnections among R&E networks with no commercial networks
- Limited colo at new HKIX1b site when available

Further Work in 2014-15

- More L2 ACL
- Advanced Route Server Software
- Portal for Participants
  - With L2 Netflow info
- Improved after-hour support
- IPv4: /23 -> /22 or /21???
- ISO27001?

Thank you!