Appropriate Layer-2 Interconnection Between IXPs

Keith Mitchell

NANOG31, San Francisco
24/25th May 2004

Layer-2 Interconnect

- Typical scenarios faced by IXP operators today:
  - ISPs conserving router ports by connecting router to IXP via own switch(es)
  - Remote ISPs connecting to IXP via 3rd-party metro/long-haul Ethernet circuits
  - ISPs using international “distance peering” services to avoid overseas router co-location overheads
    - usually Ethernet pseudo-circuit over MPLS

Layer-2 Interconnect

- Typical scenarios faced by IXP operators today:
  - ISPs connecting hybrid layer-2/3 bridge/router devices
  - Layer 2 backhaul from regional IXP struggling for critical mass
  - Increasing use of Ethernet as a circuit-switched SONET substitute
  - Increased competition and diversity in IXP marketplace leading to multiple IXPs per metro area

Layer-2 Interconnect Perspective

- All the above have led to pressure to interconnect different operators' infrastructure using layer-2 bridging, instead of layer-3 IP/BGP routing
- There are strong economic arguments for doing this in today's market conditions, but how appropriate is this from a technical perspective ?

Problems with Layer-2 Interconnect

- Fault detection difficult
  - Can only see end-to-end or nearest hop failure
  - No information on state of intermediate hops via e.g. IP traceroute or SONET loop-back
  - No visibility into networks of intermediate party(s)
- Loops and broadcast storms can impact more than one operator

Problems with Layer-2 Interconnect

- Limitations of 802.1 Spanning Tree
  - Limited or no support for multiple routing domains
  - Makes diversity protection of inter-operator links very difficult
  - Risks of topology disruption from “BPDU leaks” (e.g. 6th May ☹)
- Scaling Issues
  - VLAN tag space limited (12-bit) and not globally unique
  - Non-unicast traffic
  - Tracking/filtering legal MAC addresses

Co-Terminous IXPs

- Competition has its advantages, but too many IXP operators in the same region can *increase* participant ISPs' costs and decrease IXP viability by splitting critical mass
- *Co-Terminous* IXPs are defined here as those which share one or more buildings in the same metro area
- i.e. they can be interconnected via:
  - native, wire-speed media
  - relatively cheaply (minimal additional active components)
  - usually in-building cross-connect
- Exact $/€ and mile/km values of "cheap" and "metro area" will vary depending on local conditions

Co-Terminous Interconnect Advantages

- Reduces number and cost of connections for peering participants
- Increases “critical mass” for interconnecting IXPs
- Reduces latency and IP hop-count for traffic between participants
- Increases localization of traffic within area
- Simplifies IXP selection decision for potential participants

Issues with Non Co-Terminous Layer- 2 Interconnect

- All the issues outlined above, plus some specific additional ones:
- There may be more than one intermediate party
  - exacerbates problem diagnosis and fault finding issues
  - makes it harder to prevent and detect “dangerous” traffic
- Long-haul Ethernet circuits will likely be less transparent than native or IP-only circuits
  - latency, MTU size, traffic shaping
- Provisioning may require tracking many non-globally unique circuit identifiers (e.g. VLAN tags)

Broadcast Traffic

- This is particularly problematic in this environment
- Typical broadcasts at even large IXPs do not exceed ~100pps in normal operation
- Today's switches can forward broadcast storms at much higher rates (e.g. 10,000pps)
- But most routers connected to shared peering LANs exhaust CPU resources long before this
  - impacts many participants
- Many-to-many layer-2 interconnection between switch fabrics both increases risks and impairs scalability

XchangePoint and LoNAP

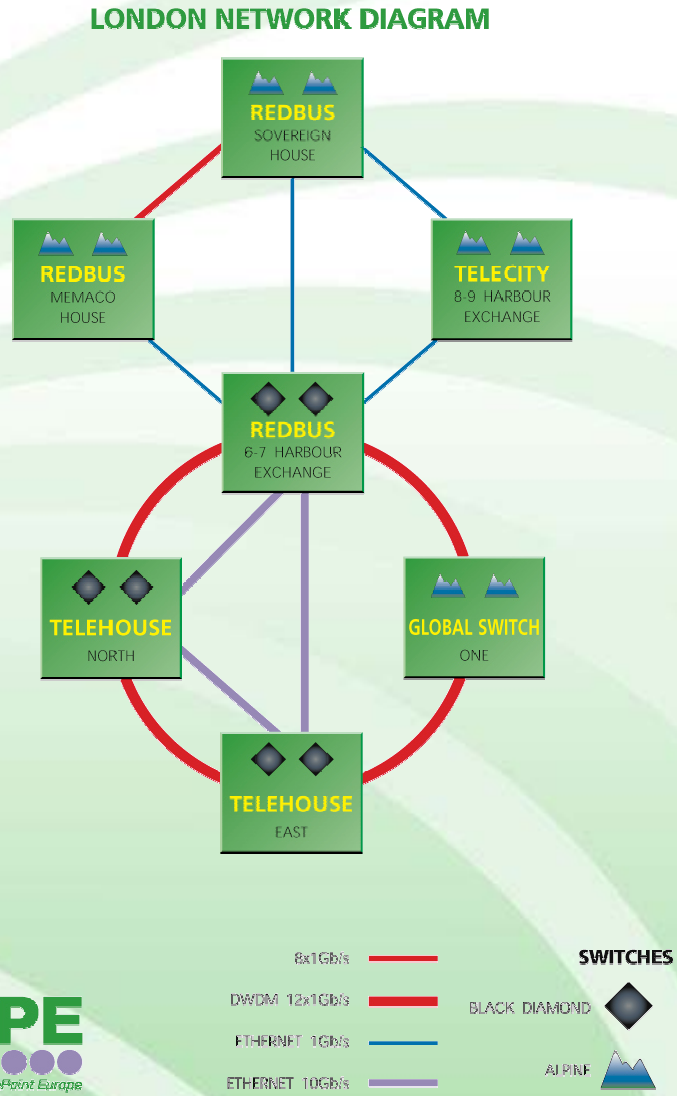
- Both London-based IXP operators, but with diverse approaches to overlapping markets
- Both competing with dominant incumbent IXP operator in London
- Informal co-operation since Q2 2002
- Formal Interconnect Agreement signed during Jan 2003
- http://www.xchangepoint.net/ourpartners/LoNAP-XP-iconnect.html

XchangePoint and LoNAP

- XchangePoint:
  - commercial IXP operator
  - established 2000
  - 150 customers, ~8Gb/s total traffic
  - transit, peering, and DSL interconnect backed by SLA
- LoNAP:
  - not-for-profit membership organization
  - established 1997
  - 40 members, ~200Mb/s traffic
  - volunteer best efforts peering

XchangePoint Network

- London
  - 7 buildings at 6 co-lo providers
  - 3 in common with LoNAP
- Frankfurt
  - 5 buildings at 3 co-lo providers
- Hamburg
  - 1 site
- Amsterdam
  - 2 sites live June 04
- *Connections **within**, not between, each metro area*

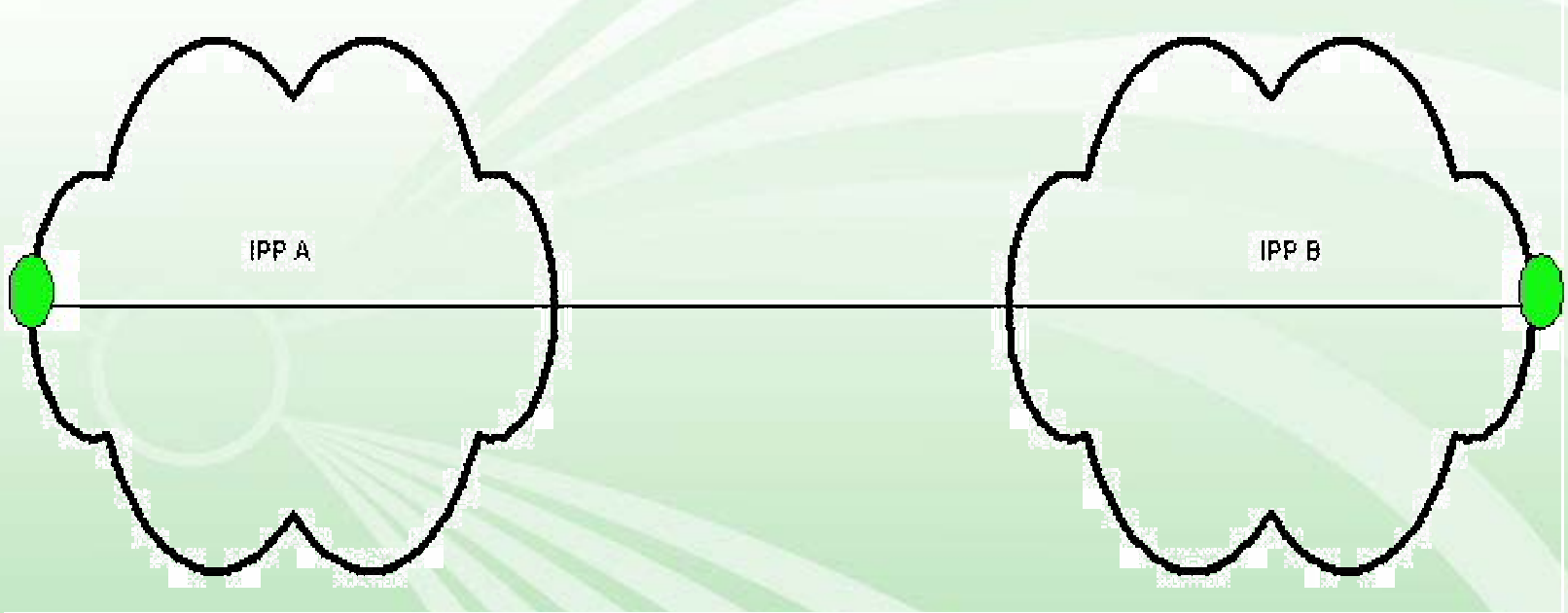


Interconnect Modes

- Identified various, mostly VLAN-based, approaches
- In practice two of these have been implemented:
  - **Mode 1:** Private Peering
  - **Mode 2**: Shared Public Peering
- Agreed to consider these and other options in future, but approach is one step at a time
  - Minimize operational risks
  - Build confidence, particularly that IXP operators would not cannibalize each others' revenues

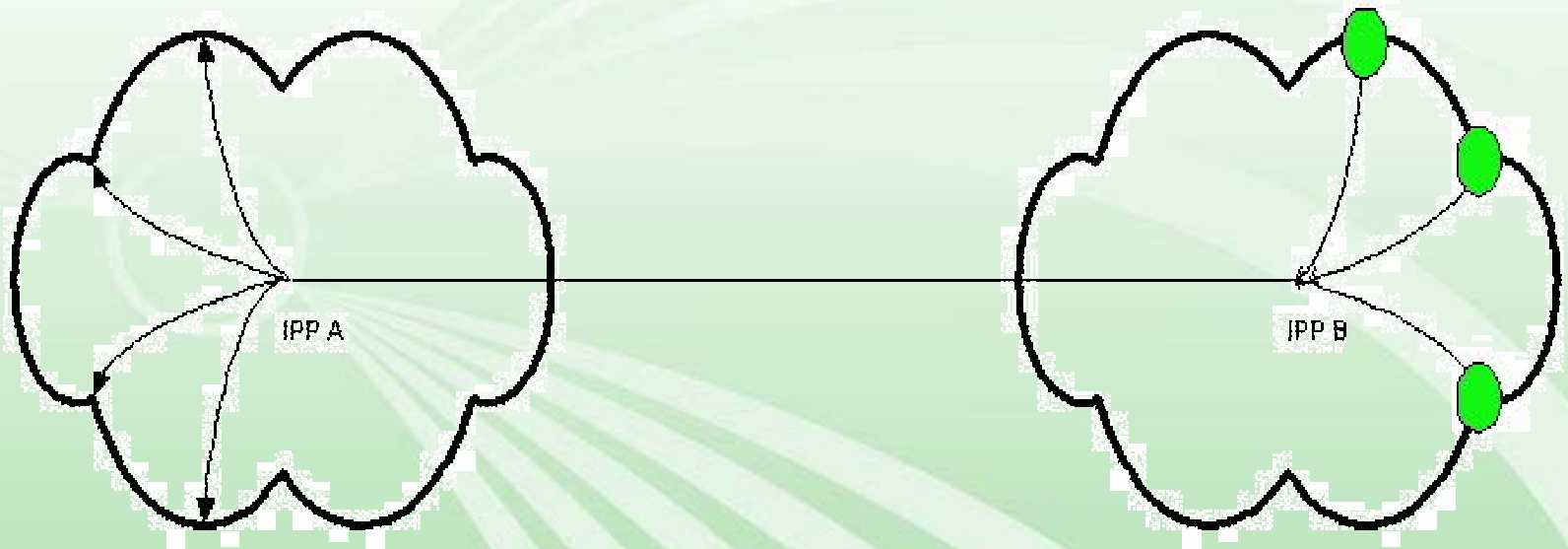
Mode 1: Private Peering across Interconnect

- Participant on IXP **A** can use point-to-point VLAN to peer with participant on IXP **B**



Mode 2: Public Peering across Interconnect

- Creates VLAN which is single logical shared public peering fabric across two physical exchange
- Participants of either IXP can opt-in (using 802.1q sub-interfaces) to this VLAN to peer with participants of other IXP



Interconnect Status

- Point-to-point private VLANs:
  - 12 XPE customers, 8 LoNAP members, 28 VLANs
  - About 8Mb/s traffic
  - Point-to-point IP address assignment is peers' responsibility
  - Not much growth since public interconnect introduced
- Shared public interconnect:
  - Introduced September 2003
  - 42 XPE customers, 14 LoNAP members
  - About 30Mb/s traffic
  - 195.47.243/24, VLAN 550

Use of VLANs

- VLANs are generally very effective at containing (e.g. broadcast) problems
- Have assigned block of VLAN tags which are unique to both IXPs
- These also need to be unique across any other layer-2 operators interconnecting with either
- 12-bit address space for unique IDs is not large !
- Block of 100 tags assigned across interconnect:
  - 40 LoNAP
  - 40 XchangePoint
  - 20 in middle public/reserved

Commercial Model

- Major principles:
  - No settlement between operators for traffic across Interconnect
  - LoNAP Members do not pay XchangePoint for use of interconnect
  - XchangePoint Customers do not pay LoNAP for use of interconnect

Commercial Model

- Commercial arrangements, e.g. peering, transit are a bi-lateral matter for participants
- Either operator has right to define own commercial terms on own participants for VLAN participation
- Above simplifies formal relationship while preserving autonomy of both IXP operators
- Other commercial models possible (e.g. revenue sharing, re-sale), but not appropriate for this relationship

Resilience

- Spanning Tree is not really practical between two different operators' layer-2 networks
  - 802.1s may change this in future
- STP traffic prohibited across interconnect
- Basic resilience implemented by multiple links, however:
  - in different locations (Telehouse East, Redbus)
  - different links for different interconnect modes
  - use manual configuration to ensure only one link per mode up at one time
- Participants wanting higher resilience should connect to **both** IXP operators !

Acceptable Use

- Very simple approach
- Any traffic traversing interconnect must conform to AUP/rules of **both** IXPs
- All traffic across interconnect **must** have explicit (non-default) VLAN tag from permitted range
- Obligations upon both operators to:
  - make all participants aware of above and changes
  - notify all affected parties in the event of any breaches
- Right to suspend interconnect mode(s) in the event of persistent unresolved breach

Service Levels

- A given operator's service level responsibility covers their own infrastructure only, and does not extend across the interconnect
- Operators must provide each other with 24x7 contact points
- Participants should send support requests to their own operator, and copy other operator
- Each operator should raise faults across the interconnect with the other operator
- Obligation on operators to inform other of outages, maintenance etc.

Documentation

- Updateable schedule to agreement sets out:
  - Physical demarcation points and ports
  - Address ranges (IPv4 and IPv6)
  - VLAN tag assignments
  - Contact points
- Web pages accessible to all participants lists:
  - VLAN assignments
  - Names, AS numbers, IP addresses of participants

Administrative Considerations

- Extent to which agreement is legally binding
- Termination notice period
- Review points defined by duration and/or traffic volumes
- Collector routers on both side of public interconnect

Open Issues

- Broadcast storms can still get across the interconnect, but usually only affect shared VLAN and/or mutual participants
- Managing and synchronizing mailing list open to interconnect participants on both sides for relaying peering requests
- Some switch vendors' use of default VLAN 1 problematic
- Using one operator to extend geographic reach of another
- What are benefits for ISPs of connecting to both ?

Observations

- Quite a lot of suspicion between operators at inception of agreement
  - Membership distrust of commercial operators' motives
  - Commercial concerns about loss of revenue
- Both IXP communities now agree there are significant mutual benefits
- It would be *very* easy to come up with a much more complex agreement
- Some additional switch/router vendor features would make life easier….

Wish List (1)

- Better protection against broadcast storms
  - Block all non-ARP broadcast packets
  - Fine granularity rate-limiting of broadcast packets (e.g. <100 pps)
  - Filter ARP packets by IP address range
- Non-STP loop detection and prevention
  - Block/ignore/reject alien BPDUs
- Ability to monitor and diagnose intermediate layer-2 hops
  - e.g. MARP (draft-retana-marp-03.txt)
  - IP-aware network probes ?

Wish List (2)

- Better choice of entry-level BGP-capable routers
- VLAN tools
  - tag re-mapping
  - larger number space ?
  - global public mapping registry ?
- Distance peering offerings which perform local spoofing of ARP broadcasts

Some Conclusions

- Layer-2 interconnect can be valid where it makes the Internet scale better
- Appropriate bi-lateral metro-area layer 2 interconnect between *co-terminous* IXPs can lead to a cheaper and simpler Internet
- Layer-2 interconnect via too many provider, switch or km hops leads to a cheaper, less stable, and more complex Internet...

Contact Details

Keith Mitchell
www.xchangepoint.net
keith@xchangepoint.net
+44 20 7395 6020

James Rice
www.lonap.net
james_r@jump.org.uk

Presentations:

- http://www.xchangepoint.net/info/nanog31-xpe-lonap.pdf
- http://www.xchangepoint.net/info/Xchange-LoNAP.ppt
- http://www.xchangepoint.net/info/IPP-interconnect.ppt