

# ISP Peering & Transit Network Design

## ISP Workshops



These materials are licensed under the Creative Commons Attribution-NonCommercial 4.0 International license (<http://creativecommons.org/licenses/by-nc/4.0/>)

# Acknowledgements

---

- This material originated from the Cisco ISP/IXP Workshop Programme developed by Philip Smith & Barry Greene
  - I'd like to acknowledge the input from many network operators in the ongoing development of these slides, especially Mark Tinka of SEACOM for his contributions
- Use of these materials is encouraged as long as the source is fully acknowledged and this notice remains in place
- Bug fixes and improvements are welcomed
  - Please email *workshop (at) bgp4all.com*

Philip Smith



# ISP Network Design

---

- Goals
- Peering
- Upstream Connectivity
- Case Study

# Goals



What does a network operator need to achieve today?



# Network Operator Goals?

---

- Today, the vast majority of content consumed by end-users is available by peering:
  - The major content providers (Google, Facebook, etc)
  - Private cross connects
  - Internet Exchange Points
- A network operator's goal is to obtain as much peering as possible
- Transit is for the last resort, for any content not available by peering

# Network Operator Goals?

---

## □ Peering

- Locally with direct cross-connect with other providers
- Locally at an Internet Exchange Point
- Getting to the nearest IXP or other interconnect

## □ Transit

- Relying on another network operator to get the rest of the Internet
- Considered a last resort now

# Peering



Interconnecting networks

# Peers

---

- A peer is another autonomous system with which the local network has agreed to exchange locally sourced routes and traffic
- Private peer
  - Private link between two providers for the purpose of interconnecting
- Public peer
  - Internet Exchange Point, where providers meet and freely decide who they will interconnect with
- **Recommendation: peer as much as possible!**

# Common Mistakes

---

- ❑ Mistaking a transit provider's for profit "Exchange" business for a no-cost public peering point
- ❑ Not working hard to get as much peering as possible
  - Physically near a peering point (IXP) but not present at it
  - (Transit is rarely cheaper than peering!!)
- ❑ Ignoring/avoiding competitors because they are competition
  - Even though potentially valuable peering partner to give customers a better experience

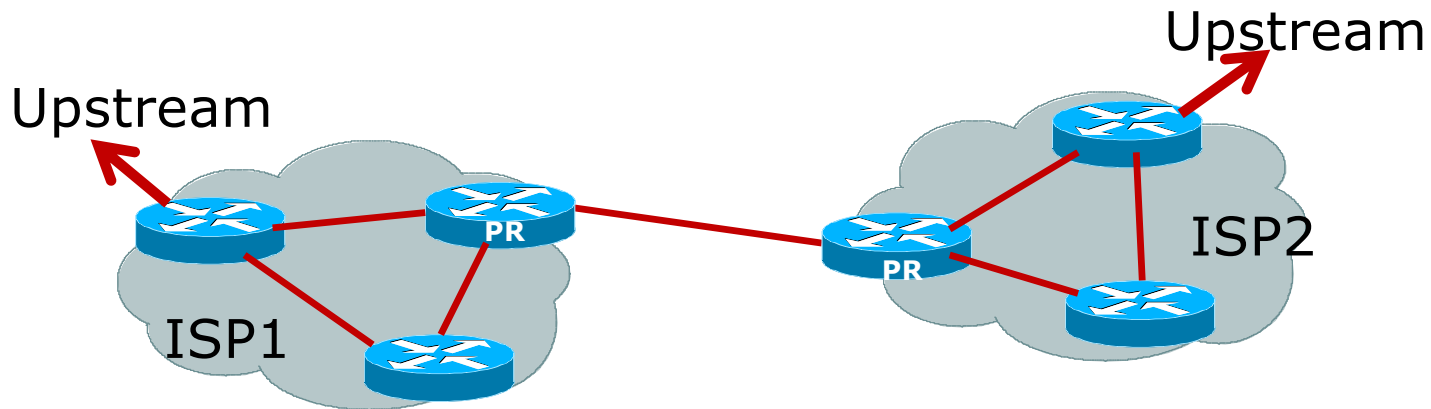
# Private Interconnection: What it is

---

- Two service providers agree to interconnect their networks
  - They exchange prefixes they originate into the routing system (usually their aggregated address blocks)
  - They share the cost of the infrastructure to interconnect
    - Typically each paying half the cost of the link (be it circuit, satellite, microwave, fibre,...)
    - Connected to their respective peering routers
  - Peering routers only carry domestic prefixes

# Private Interconnection: Detail

---



- PR = peering router
  - Runs iBGP (internal) and eBGP (with peer)
  - No default route
  - No “full BGP table”
  - Domestic prefixes only
- Peering router used for all private interconnects

# Private Interconnect: Where?

---

- Private Interconnects can be established anywhere
  - Where two providers are in the same facility
    - Usually simple fibre cross-connect between two peering routers
    - Most common scenario – datacentres, at IXP facilities, etc
  - Between two providers with PoPs in the same metro area
    - Will involve obtaining and sharing the costs of installing fibre (or other media) between the two locations
    - The more traditional/historical type of interconnect



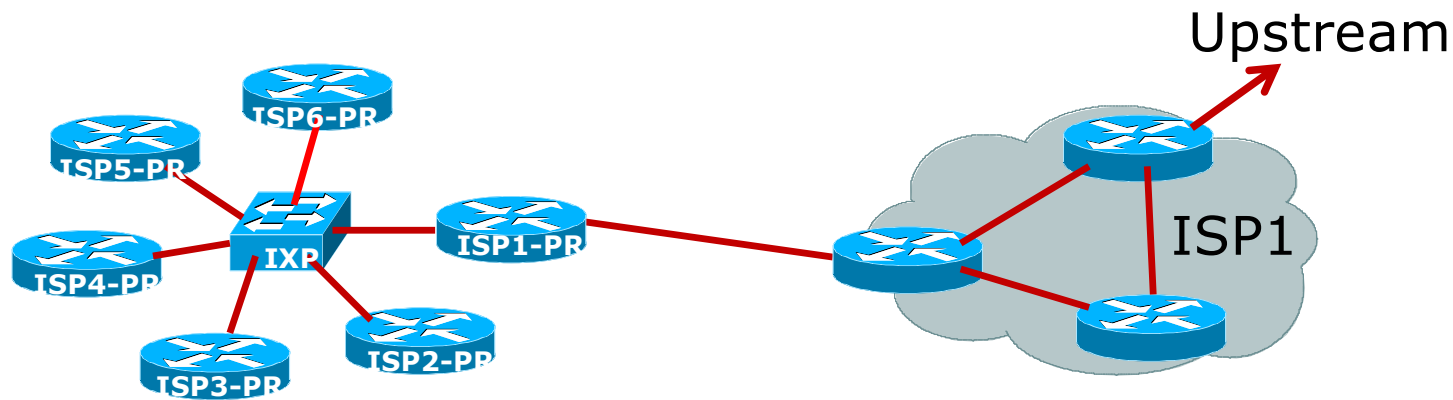
# Public Interconnection: What it is

---

- Service provider participates in an Internet Exchange Point
  - It exchanges prefixes it originates into the routing system with the participants of the IXP
  - It chooses who to peer with at the IXP
    - Bi-lateral peering (like private interconnect)
    - Multi-lateral peering (via IXP's route server)
  - It provides the router at the IXP and provides the connectivity from their PoP to the IXP
  - Their IXP router carries only the prefixes they will share with other peers across the IXP

# Public Interconnection: Detail

---



- ISP1-PR = peering router of our ISP
  - Runs iBGP (internal) and eBGP (with IXP peers)
  - No default route
  - No “full BGP table”
  - Domestic prefixes only
- Usually physically located at the IXP

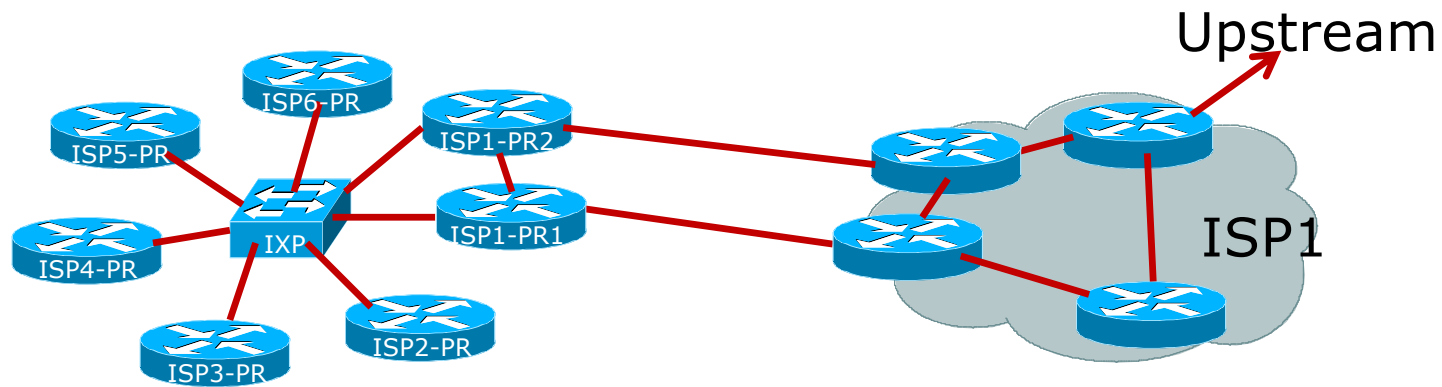
# Public Interconnection

---

- The ISP's router IXP peering router needs careful configuration:
  - It is remote from the domestic backbone
  - Should not originate any domestic prefixes
  - (As well as no default route, no full BGP table)
  - Filtering of BGP announcements from IXP peers (in and out)
  
- Provision of a second link to the IXP:
  - (for redundancy or extra capacity)
  - Usually means installing a second router
    - Connected to a second switch (if the IXP has two more more switches)
    - Interconnected with the original router (and part of iBGP mesh)

# Public Interconnection

---



- Provision of a second link to the IXP means considering redundancy in the SP's backbone
  - Two routers
  - Two independent links
  - Separate switches (if IXP has two or more switches)

# What if there is no local IXP?

---

- ❑ If there is no local IXP, one is usually created by the network operators once there are more than two who wish to interconnect
- ❑ Private peering means that the three operators have to buy circuits between each other
  - Works for three operators, but adding a fourth or a fifth means this does not scale
- ❑ Solution:
  - Internet Exchange Point

# Internet Exchange Point

---

- Every participant has to deploy just one link
  - From their premises to the IXP
- Rather than N-1 links to connect to the N-1 other ISPs
  - 5 ISPs will have to share the cost of 4 links = 2 whole links → already twice the cost of the IXP connection
- Today metro area connectivity to get to a local IXP is easy using fibre-optics
  - Which means 10Gbps speeds is inexpensive to do
  - Most IXP switch ports now start at 10Gbps (and offer 1Gbps for smaller operators)

# Internet Exchange Point

---

## □ Solution

- Every operator participates in the IXP
- Cost is minimal – one local link covers all domestic traffic
- International links are used for just international traffic – and backing up domestic links in case the IXP suffers any outage

## □ Result:

- Local traffic stays local
- QoS considerations for local traffic is not an issue
- RTTs between members are typically sub 1ms
- Customers enjoy the Internet experience
- Local Internet economy grows rapidly

# Who can join an IXP?

---

- Requirements are very simple: any organisation which operates their own autonomous network, and has:
  - Their own IP address space
  - Their own AS number
  - Their own transit arrangements
- This often includes:
  - Commercial ISPs
  - Academic & Research networks
  - Internet infrastructure operators (eg Root/ccTLDs)
  - Content Providers & Content Distribution Services
  - Broadcasters and media
  - Government Information networks

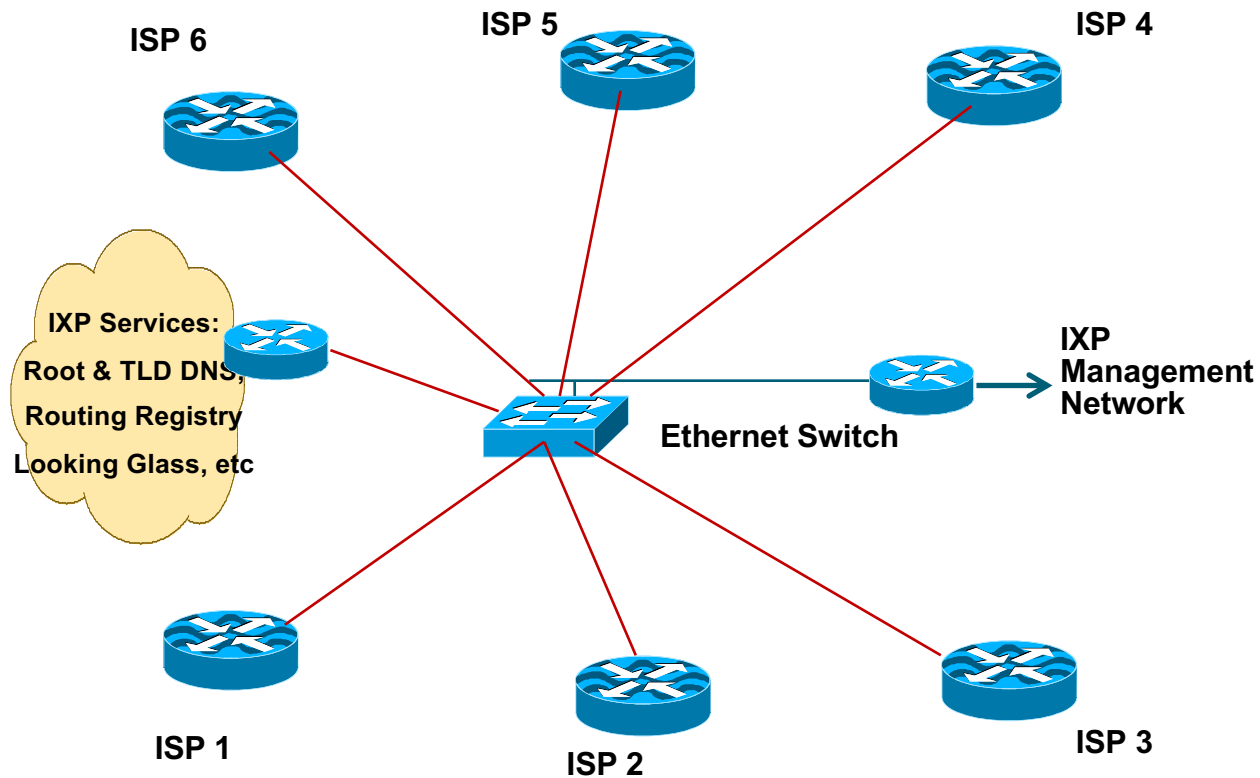


# IXP Design

---

- Very simple concept:
  - Ethernet switch is the interconnection media
    - IXP is one LAN
  - Each ISP brings a router, connects it to the ethernet switch provided at the IXP
  - Each ISP peers with other participants at the IXP using BGP
- Scaling this simple concept is the challenge for the larger IXPs
- Known as a Layer-2 Exchange Point

# Internet Exchange Point



Single site internet exchange point

# IXP Features

---

- Neutral location
  - Anyone can install fibre or other connectivity media to access the IXP
    - Without extra cost or regulations imposed by location
- Secure location
  - Thorough security, like any other network data centre
- Accessible location
  - Easy/convenient for all participants to access
- Expandable location
  - IXPs result in Internet growth, and increasing space requirements within the facility

# IXP Features

---

- Operation:
  - Requires neutral IXP management
  - “Consortium”
    - Representing all participants
    - “Management Board” etc
- Funding:
  - All costs agreed and covered equally by IXP participants
  - Hosting location often contributes – the IXP brings them more business
- Availability:
  - 24x7 cover provided by hosting location
    - Managed by the consortium

# IXP Standards

---

- Industry Standards documented by Euro-IX, the European IXP Association
  - Contributed to by the Euro-IX members
  - <https://www.euro-ix.net/en/forixps/set-ixp/>
- IXP BCP
  - General overview of the infrastructure, operations, policies and management of the IXP
  - <https://www.euro-ix.net/en/forixps/set-ixp/ixp-bcops/>
- IXP Website BCP
  - <https://www.euro-ix.net/en/forixps/set-ixp/ixp-bcops/ixp-website/>

# Services Offered by IXPs

---

- Root server
  - Anycast instances of F, I and L root nameservers are present at many IXEs
- ccTLD DNS
  - The country IXP could host the country's top level DNS
  - e.g. "SE." TLD is hosted at Netnod IXEs in Sweden
  - Offer back up of other country ccTLD DNS
- gTLD DNS
  - .com & .net are provided by Verisign at many IXEs

# Services Offered by IXPs

---

## □ Route Server

- Helps scale IXes by providing easier BGP configuration & operation for participants with Open Peering policies
- Technical detail covered later on

## □ Looking Glass

- One way of making the Route Server routes available for global view (e.g. [www.traceroute.org](http://www.traceroute.org))
- Public or members-only access

# Services Offered by IXPs

---

- Content Redistribution/Caching
  - Various providers offering content distribution services
  - Broadcast media
- Network Time Protocol
  - Locate a stratum 1 time source (GPS receiver, atomic clock, etc) at IXP
- Routing Registry
  - Used to register the routing policy of the IXP membership (more later)



# Notes on IXP Services

---

- If IXP is offering services to members:
  - Services need transit access
  - Transit needs to be arranged with one or two IXP members (cost shared amongst all members)
  
- Consider carefully:
  - Should services be located at the IXP itself?
    - How to arrange and pay for the transit to those services?
  - or-
  - Should services be hosted by members and shared with the others?

# What if there is no local IXP?

---

- If there is no local IXP, and there aren't sufficient operators to justify creating one:
  - Private Network Interconnect with other operator
  - Purchase capacity (bandwidth) to get to the topologically closest major interconnect (RTT matters!)
- Many major locations around the world are focal points of operator interconnects
  - These are known as Regional IXPs

# Regional Internet Exchange Point

---

- These are also “local” Internet Exchange Points
- But also attract regional ISPs and ISPs from outside the locality
  - Regional ISPs peer with each other
  - And show up at several of these Regional IXPs
- Local ISPs peer with ISPs from outside the locality
  - They don't compete in each other's markets
  - Local ISPs don't have to pay transit costs
  - ISPs from outside the locality don't have to pay transit costs
  - Quite often ISPs of disparate sizes and influences will happily peer – to defray transit costs

# Examples of Regional IXPs

---

- Sydney
  - Serves Australia, NZ and much of the Southern Pacific
- Singapore
  - Serves South & South East Asia
- Hong Kong
  - Serves South East Asia
- Tokyo
  - Serves East & South East Asia
- London/Amsterdam/Frankfurt
  - Serve Europe, Africa, Middle East
- Los Angeles, Bay Area, Seattle
  - Serve Asia, Pacific and North America
- New York, Washington, Miami
  - Serve Europe & Latin America

All attract operators from all around the world

All encourage interconnection

# What should operators do?

---

- Many operators participate in their local IXP
  - Keeps local traffic local
  - Gives best experience to the end-user for content
  
- Many operators also purchase connectivity (bandwidth) to Regional IXPs
  - Bandwidth as IPLC (international private leased circuit)
    - **NOT** buying transit to the Regional IXP
  - And establish peering across the IX fabric
  - And establish PNI with major content operators for Cache fill

## Footnote: "Layer 3 IXPs"

---

- Some entities talk about Layer 3 Internet Exchange Points
  - These are not IXPs
- Layer 3 IXP today is marketing concept used by Transit ISPs
  - Some incumbent telecom operators call their domestic or international transit businesses "Exchanges"
- Real Internet Exchange Points are only Layer 2
  - L2 is the accepted International standard

## “Layer 3 IXP” – what breaks

---

- One extra AS hop between peers
  - Makes path via IXP suboptimal/less preferred
  - Path between peers usually remains with upstream transit provider
    - Unless both peers actively implement BGP policies to prefer the L3 IXP
- Members cannot peer with whom they please
  - Mandatory multilateral peering
  - Third party (L3 IXP operator) required to configure peering sessions and peering policy

## “Layer 3 IXP” – what breaks

---

- More complicated troubleshooting
  - Troubleshooting peering problems has to involve IXP operator too
- No policy control
  - BGP attributes shared between members get dropped by IXP router
  - (Examples are BGP communities, MEDs)



## “Layer 3 IXP” – what breaks

---

- CDNs won't join
  - They have requirements to peer directly with IXP members
- Redundancy problems
  - L3 IXPs with dual sites appear as two separate transit providers between peers
  - Traffic engineering?
- L3 “IXP” Operator requires strong BGP skills

# Upstream Connectivity



# Transits

---

- Transit provider is another autonomous system which is used to provide the local network with access to other networks
- Access for
  - Local traffic only
  - Maybe local and regional traffic
  - Content Cache fill for a locally hosted Cache
  - But more usually the whole Internet

# Transits

---

- Transit providers need to be chosen wisely:
  - Only one
    - No redundancy
  - Too many
    - Very difficult to load balance
    - No economy of scale (costs more per Mbps)
    - Hard to provide good service quality
- **Recommendation: at least two, no more than three**

# Common Mistakes

---

- Operators sign up with too many transit providers
  - Results in lots of small circuits (cost more per Mbps than larger ones)
  - Transit rates per Mbps reduce with increasing transit bandwidth purchased
  - Hard to implement reliable traffic engineering that doesn't need daily fine tuning depending on customer activities
  
- No diversity
  - Chosen transit providers all reached over same satellite or same submarine cable
  - Chosen transit providers themselves have poor onward transit and peering arrangements

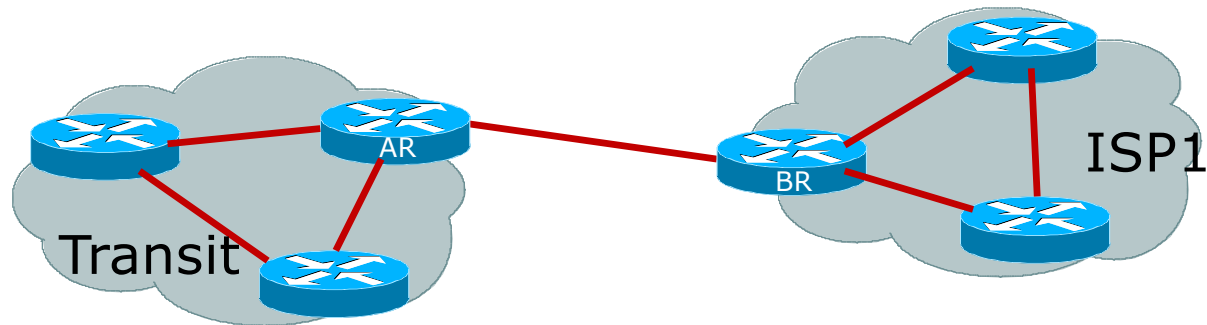
# Upstream/Transit Connection

---

- Two scenarios:
  - Transit provider is in the locality
    - Which means bandwidth is cheap, plentiful, easy to provision, and easily upgraded
  - Transit provider is a long distance away
    - Over undersea cable, satellite, long-haul cross country fibre, etc
- Each scenario has different considerations which need to be accounted for

# Local Transit Provider

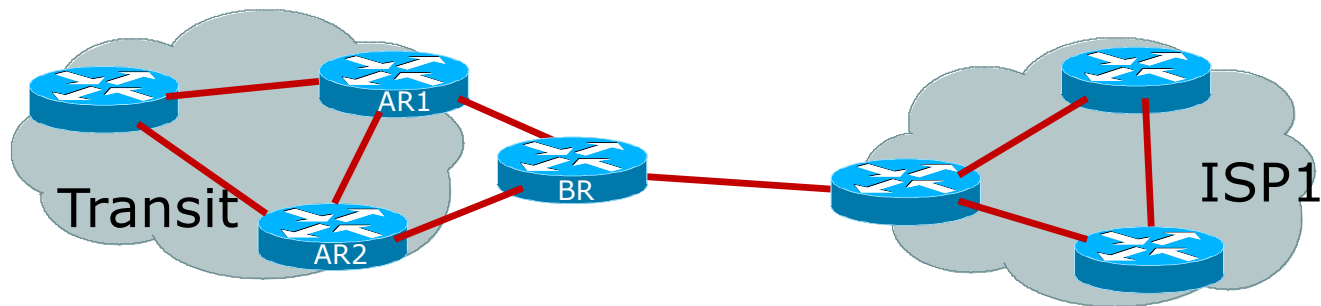
---



- BR = ISP's Border Router
  - Runs iBGP (internal) and eBGP (with transit)
  - Either receives default route or the full BGP table from upstream
  - BGP policies are implemented here (depending on connectivity)
  - Packet filtering is implemented here (as required)

# Distant Transit Provider

---



- BR = ISP's Border Router
  - Co-located in a co-lo centre (typical) or in the upstream provider's premises
  - Runs iBGP with rest of ISP1 backbone
  - Runs eBGP with transit provider router(s)
  - Implements BGP policies, packet filtering, etc
  - Does not originate any domestic prefixes



# Distant Transit Provider

---

- Positioning a router close to the Transit Provider's infrastructure is strongly encouraged:
  - Long haul circuits are expensive, so the router allows the ISP to implement appropriate policies first
  - Moves packet buffering away from the Transit provider
    - Their router may not have the packet buffer sizing to support long haul links
  - Using remote co-lo allows the ISP to choose another transit provider and migrate connections with minimum downtime

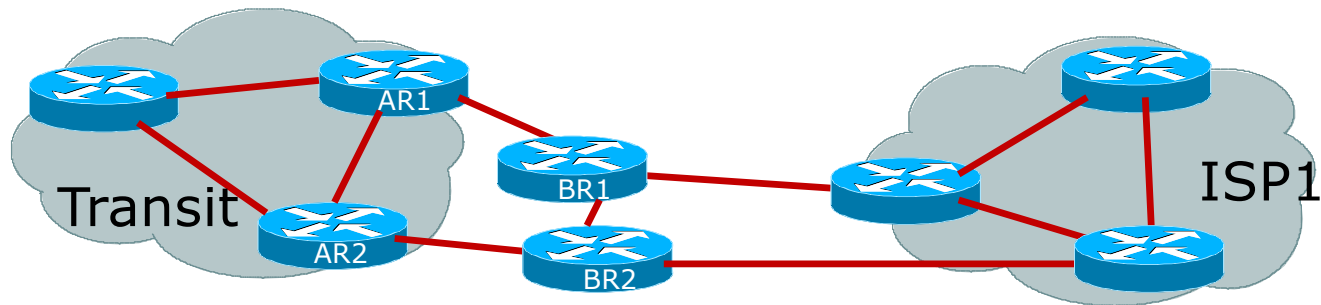
# Distant Transit Provider

---

- Other points to consider:
  - Does require remote hands support
  - (Remote hands would plug or unplug cables, power cycle equipment, replace equipment, etc as instructed)
  - Appropriate support contract from equipment vendor(s)
  - Sensible to consider two routers and two long-haul links for redundancy

# Distant Transit Provider

---



- Upgrade scenario:
  - Provision two routers
  - Two independent circuits (check fibre path)
  - Consider second transit provider and/or turning up at an IXP

# Optimising Long Haul Links

---

## □ Strategies for choosing Transit Providers

### ■ Geographical diversity

- If one is in the East, choose the other one to be in the West
- For example, a South Pacific Network Operator would connect to Australia and to the US
- If the US link fails, there is back up via Australia – and vice-versa
- Traffic for Asia and Pacific goes via Australia; traffic for Europe and US goes via US

### ■ Cost

- Two transit providers optimises transit costs
- More providers means greater cost per Mbps and greater challenges to make traffic engineering work

# Optimising Long Haul Links

---

- Transit providers are too often focused on being a monopoly
  - Unless legislated, this is a failed strategy
  - Monopolies tend to be bypassed, and only harm the country with the monopoly
- The important criteria today are:
  - Round Trip Time (RTT) – latency
  - Bandwidth
  - Reliability
- Every network operator goal needs to be to minimise RTT for all traffic, provide at maximum bandwidth, and with maximum reliability

# Examples: Pacific

---

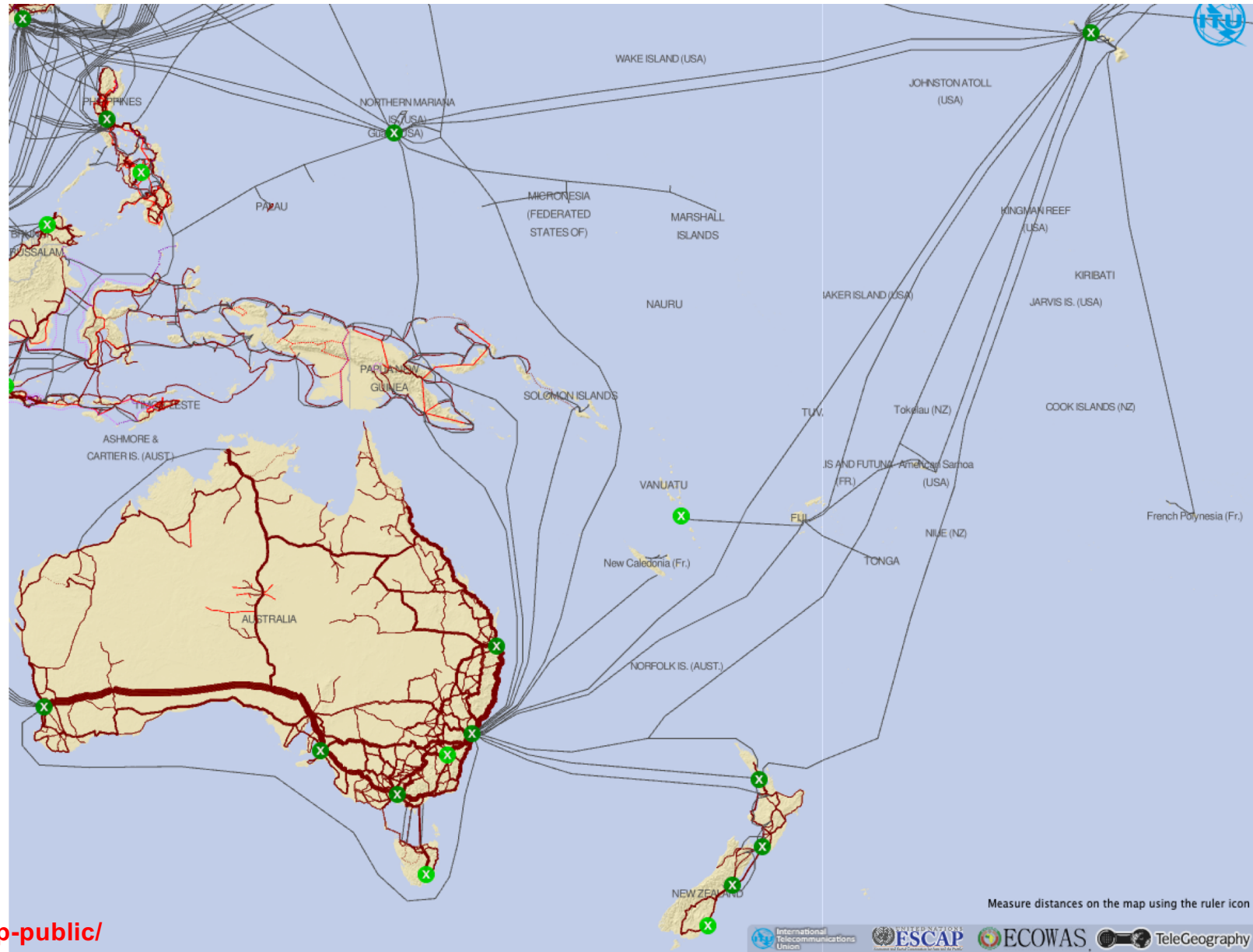
- Sydney and Los Angeles are the interconnect hubs for the Pacific
  - There are more optimum locations which offer much better RTT and performance than hauling traffic to/from/via Sydney and/or Los Angeles
- The PacPeer Project explores optimum interconnections for network operators across the Pacific
  - <https://pacpeer.org/>
  - [https://pacpeer.org/presentations/brewerj\\_peering\\_strategy\\_pacific\\_pacnog18.pdf](https://pacpeer.org/presentations/brewerj_peering_strategy_pacific_pacnog18.pdf)

# Examples: Pacific

---

- Fiji could be the regional hub for the South Pacific
- Guam could be the regional hub for the North Pacific
- Both Fiji & Guam have:
  - Large amounts of submarine fibre passing through
  - Fiji has no open neutral interconnect facility
- Hawaii should be the regional hub for the whole Pacific
  - (following the fibre paths)
  - But capacity is cheaper direct to Los Angeles (even though latency more than doubles)
    - (Pacific to Hawaii + Hawaii to Los Angeles is more expensive than Pacific to Los Angeles)

# Pacific Fibre Map



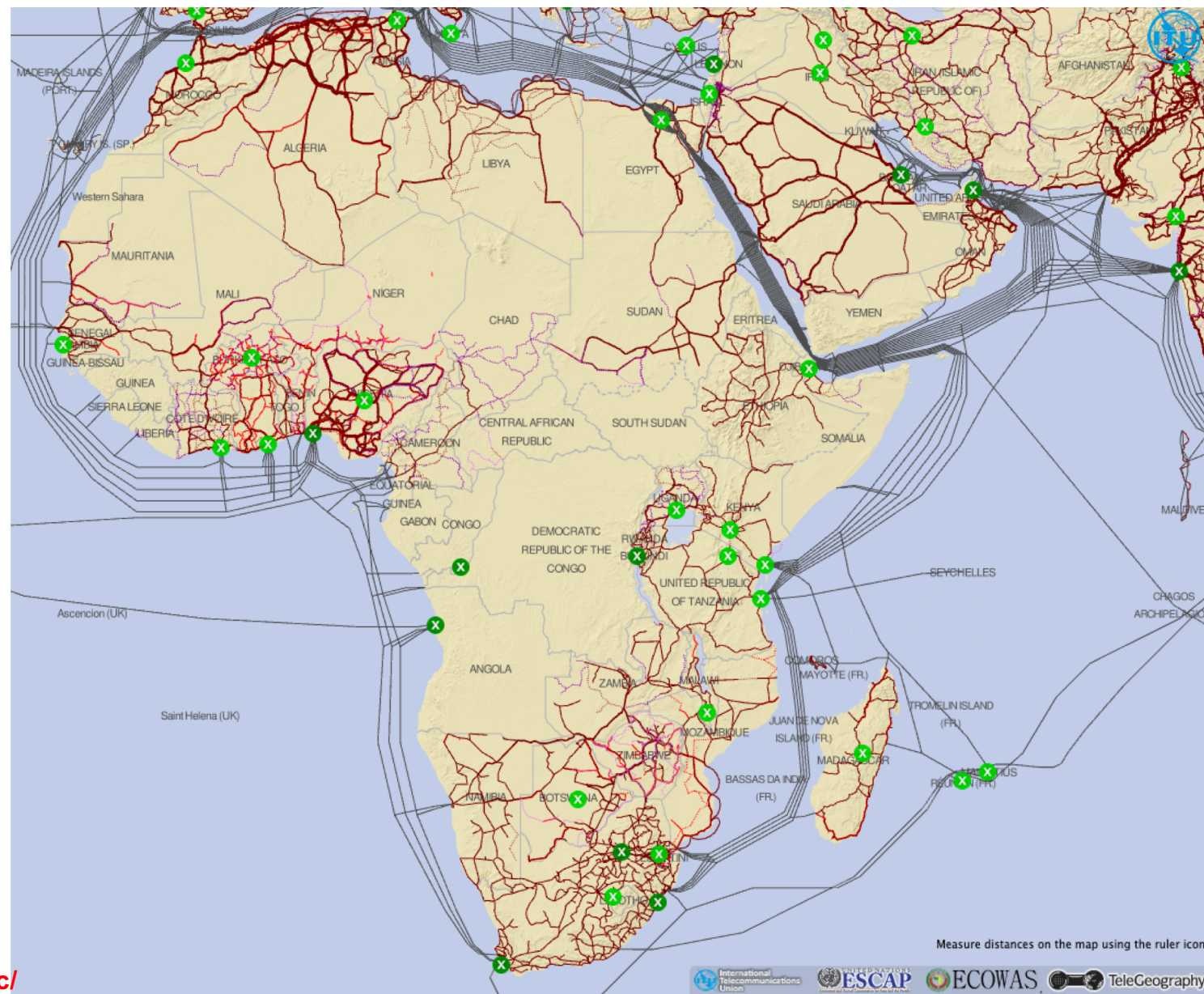


# Examples: Africa

---

- There is still no obvious regional Interconnect on the whole of the African continent
- Historically fibre went to Europe – and providers would connect based on their parent European operator
  - Inter-country traffic usually went via Europe
- Cairo, Alexandria and Djibouti could be a major hubs
  - Large amounts of fibre transit Djibouti & Egypt
  - No open neutral interconnect facility
- Mombasa (Kenya) could well become one in the near future for Eastern Africa
  - Major landing point for submarine fibre and for terrestrial fibre infrastructure
- What about Western Africa?
  - Lagos? Accra?

# Africa Fibre Map



Source:

<https://www.itu.int/itu-d/tnd-map-public/>

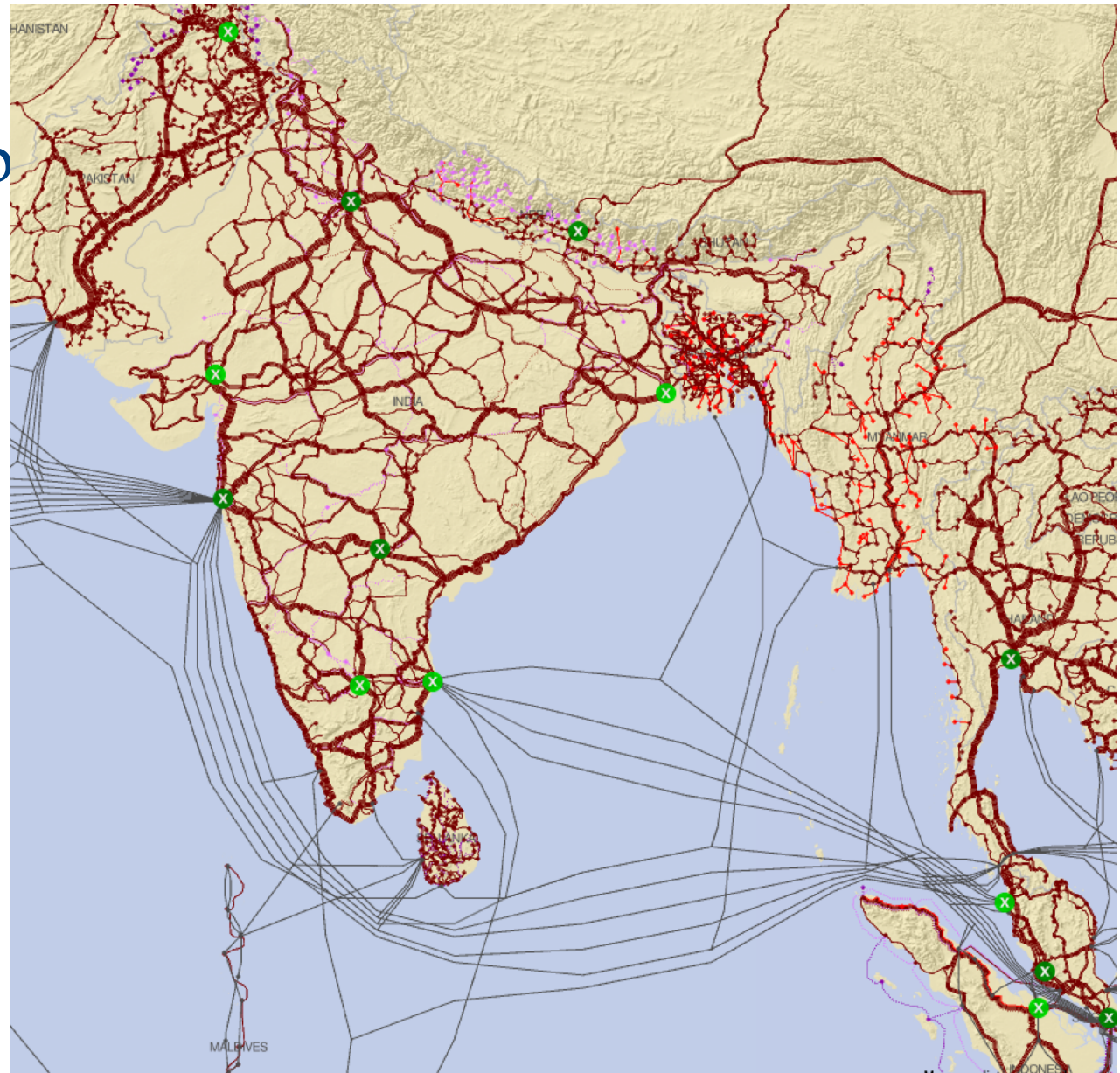
## Examples: South Asia

---

- There is still no obvious regional Interconnect in South Asia
- Mumbai and Chennai in India are obvious locations
  - Large concentrations of fibre landing in both cities
  - But no openly available transcontinental fibre interconnecting the two cities
- But only Indian licenced operators are permitted to provide transit
  - No open neutral interconnect facility
  - All traffic subject to Indian laws, even if it doesn't go to Indian consumers
  - So South Asia loses interconnect business to Singapore, which has become the interconnect for the whole region



# South Asia Fibre Map



Source:

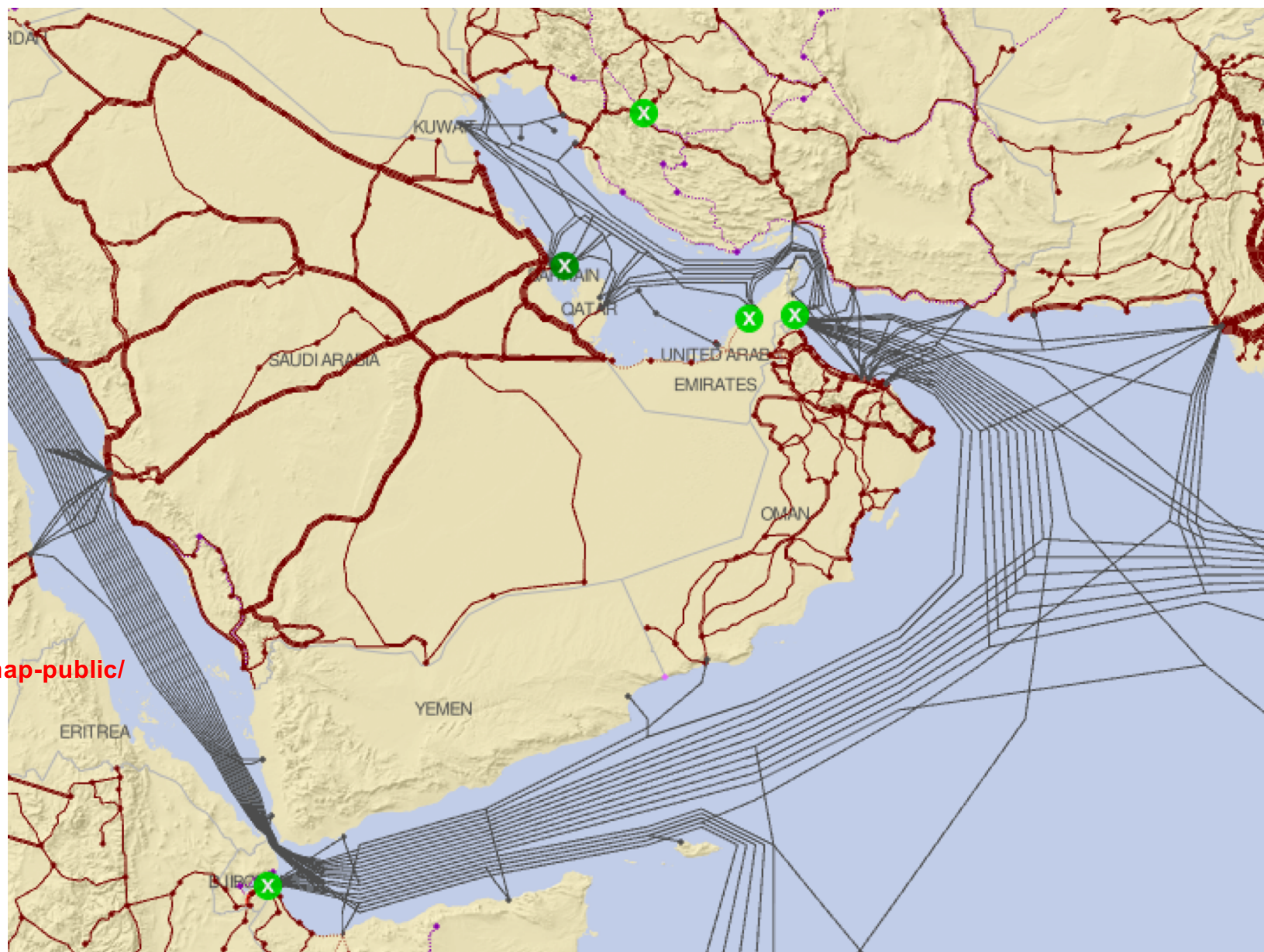
<https://www.itu.int/itu-d/tnd-map-public/>

## Examples: Middle East

---

- There is still no obvious regional Interconnect in the Middle East
- Reasons:
  - Regional rivalries, similar to those common in Asia in the 1990s
  - Everyone wants to be the hub!
- A lot of fibre lands around Fujairah (UAE)
  - Would be an obvious regional hub
- Only UAE licenced operators can provide transit & interconnects
  - No open neutral interconnect facility

# Middle East Fibre Map



Source:

<https://www.itu.int/itu-d/tnd-map-public/>



# Optimising Long Haul Links

---

- Network operators will participate in open neutral regional interconnects, where they:
  - May choose who they peer with
  - May choose who they buy transit from
  - Are not subjected to irrelevant domestic content laws
    - They are not selling services in the country in question
    - Some countries enforce domestic laws on all international transit content
- Areas without Regional Interconnects for IP traffic have no mechanisms in place to encourage these Interconnects

# Optimising Long Haul Links

---

- Summary of what's important:
  - Maximising fast and high bandwidth content delivery to end-users
  - Minimising round trip times from content to end-users
  - Enabling "next-generation" internet services
  - 5G and "Internet of Things" cannot deliver their promise using last century approach to Internet Service provision



# Upstream Connectivity and Peering Case Study



How Seacom chose their international peering locations and transit providers

# Objective

---

- Obtain high grade Internet connectivity for the wholesale market in Africa to the rest of the world
- Emphasis on:
  - Reliability
  - Interconnectivity density
  - Scalability

# Metrics Needed in Determining Solution (1)

---

- Focusing on operators that cover the destinations mostly required by Africa
  - i.e., English-speaking (Europe, North America)
- Include providers with good connectivity into South America and the Asia Pacific.
- Little need for providers who are strong in the Middle East, as demand from Africa for those regions is very, very low.

## Metrics Needed in Determining Solution (2)

---

- Split the operators between Marseille (where the SEACOM cable lands) and London (where there is good Internet density)
  - To avoid outages due to backhaul failure across Europe
  - And still maintain good access to the Internet
- Look at providers who are of similar size so as not to fidget too much (or at all) with BGP tuning.
- The providers needed to support:
  - 10Gbps ports
  - Bursting bandwidth/billing
  - Future support for 100Gbps or  $N \times 10\text{Gbps}$

## Metrics Needed in Determining Solution (3)

---

- Implement peering at major exchange points in Europe
  - To off-set long term operating costs re: upstream providers.

# Implementing Solution

---

- ❑ Connected to Level(3) and GT-T (formerly Inteliquent, formerly Tinet) in Marseille
- ❑ Connected to NTT and TeliaSonera in London
- ❑ Peered in London (LINX)
- ❑ Peered in Amsterdam (AMS-IX)
- ❑ BGP setup to prefer traffic being exchanged at LINX and AMS-IX
- ❑ BGP setup to prefer traffic over the upstreams that we could not peer away
- ❑ No additional tuning done on either peered or transit traffic, i.e., no prepending, no de-aggregation, etc. All traffic setup to flow naturally

# End Result

---

- 50% of traffic peered away in less than 2x months of peering at LINX and AMS-IX
- 50% of traffic handled by upstream providers
- Equal traffic being handled by Level(3) and GT-T in Marseille
- Equal traffic being handled by TeliaSonera and NTT in London
- Traffic distribution ratios across all the transit providers is some 1:1:0.9:0.9
- This has been steady state for the last 12x months
  - No BGP tuning has been done at all

# Design Considerations Summary





# Summary

---

- Design considerations for:
  - Private interconnects
    - Simple private peering
  - Public interconnects
    - Router co-lo at an IXP
  - Local transit provider
    - Simple upstream interconnect
  - Long distance transit provider
    - Router remote co-lo at datacentre or Transit premises

# ISP Transit & Peering Network Design



ISP Workshops