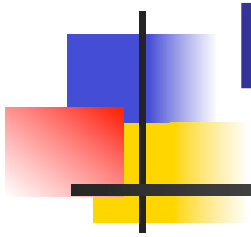


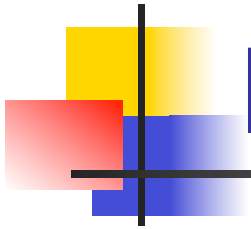
Internet Development Experiences and Lessons

Philip Smith

MENOG 13

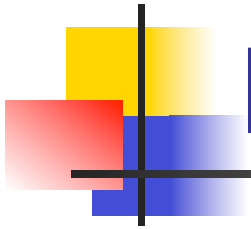
22nd September 2013

Kuwait



Background

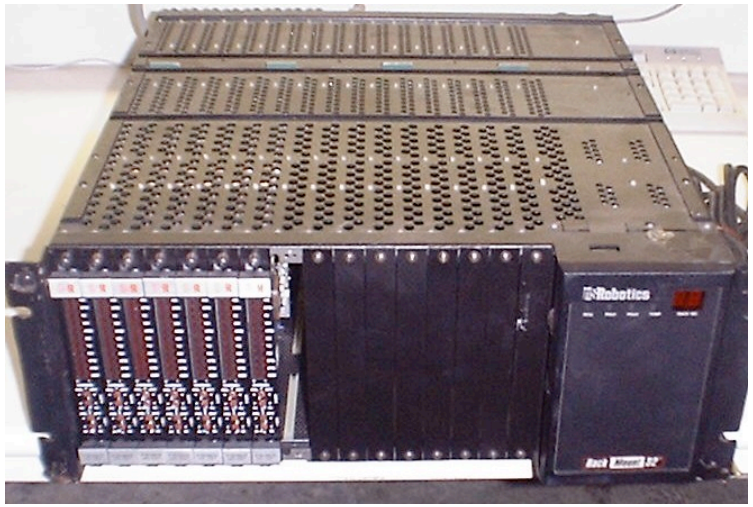
- Internet involvement started in 1989 while at University completing PhD in Physics
  - Got a little bit side-tracked by Unix, TCP/IP and ethernet
  - Helped design and roll out new TCP/IP ethernet network for Department
  - Involved in day to day operations of CAD Lab as well as Dept public Unix servers (HP and Sun)
  - Caught the Internet bug!



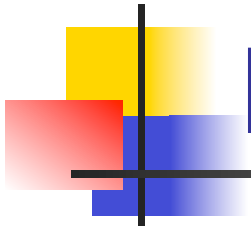
How it all started

- At end of University Post Doc in 1992
  - Job choice was lecturer or “commercial world”
  - Chose latter – job at UK’s first ISP advertised on Usenet News uk.jobs feed
  - Applied, was successful, started at PIPEX in 1993
  - First big task – upgrade modems from standalone 9.6kbps to brand new Miracom 14.4kbps rack mount
    - With upgradable FLASH for future standards upgrades!

In at the deep end

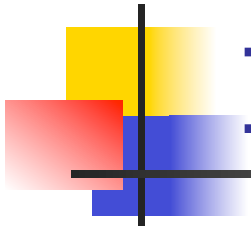


- Testing testing and more testing
- Rackmount saved space
- But did V.32bis work with all customers??



First lesson

- Apart from wishing to be back at Uni!
- Test against customers expectations and equipment too
  - Early v.32bis (14.4kbps) modems weren't always backward compatible with v.32 (9.6kbps) or older standards
  - One manufacturer's v.32bis didn't always talk to another's v.32bis – fall back to v.32 or slower
- Vendor's promises and specification sheets often didn't completely match reality

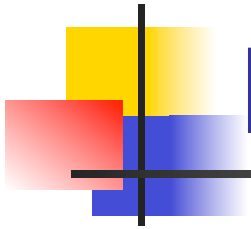


ISP Backbones

- In those early days, BGP was “only for experts”, so I watched in awe
  - Learned a little about IGRP and BGPv3
  - But not enough to be conversant
- April 1994 saw the migration from Classful to Classless BGP
  - Beta Cisco IOS had BGPv4 in it
  - Which meant that our peering with UUNET could be converted from BGPv3 to BGPv4
  - With the cheerful warning that “this could break the Internet”

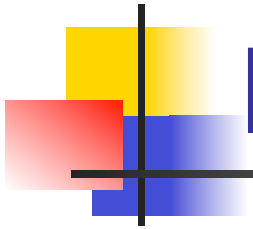
ISP Backbones

- Internet didn't break, and the whole Internet had migrated to using classless routing by end of 1994
- But classful days had left a mess behind
  - Large numbers of "Class Cs" still being announced
  - The CIDR Report was born to try and encourage these Class Cs to be aggregated
  - Cisco made lots of money upgrading existing AGS and AGS+ routers from 4Mbytes to 16Mbytes of RAM to accommodate
  - ISP engineers gained lots of scars on hands from replacing memory boards and interfaces



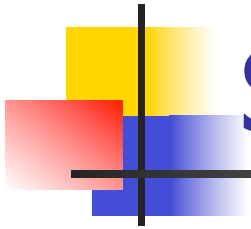
BGP improvements

- The ISP in 2013 has never had it so good!
- In 1994/5:
  - iBGP was fully meshed
  - Routers had 16Mbyte RAM
  - Customer BGP announcements only changeable during maintenance outages
  - BGP table took most of the available RAM in a router
  - The importance of separation of IGP/iBGP/eBGP was still not fully appreciated
  - No such thing as a BGP community or other labour saving configuration features



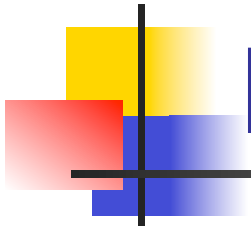
BGP improvements

- Major US ISP backbone meltdown
  - iBGP full mesh overloaded CPUs, couldn't be maintained
  - Cisco introduced BGP Confederations, and a little later Route Reflectors, into IOS
- By this point I was running our backbone operations
  - Colleague and I migrated from full mesh to per-PoP Route Reflector setup in one 2 hour maintenance window



Second Lesson

- Migrating an entire backbone of 8 PoPs and 50+ routers from one design of routing protocol to another design should not be done with out planning, testing, or phasing
  - We were lucky it all “just worked”!



Peering with the "enemy"

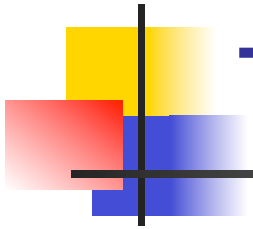
- Early PIPEX days saw us have our own paid capacity to the US
  - With a couple of paid connections to Ebone (for their "Europe" routes) and SWIPnet (as backup)
  - Paid = V Expensive
- Interconnecting with UK competition (UKnet, Demon, BTnet) seen as selling the family jewels! And would be extremely bad for sales growth
  - Even though RTT, QoS, customer complaints, extreme cost of international bandwidth, logic and commonsense said otherwise
  - But we did connect to JANET (UK academics) – because they were non-commercial and "nice guys"

Birth of LINX

- Thankfully logic, commonsense, RTT, QoS and finances prevailed over the sales fear campaign
- The technical leadership of PIPEX, UKnet, Demon, BTnet and JANET met and agreed an IXP was needed
  - Sweden had already got Europe's first IX, the SE-GIX, and that worked v nicely
- Of course, each ISP wanted to host the IX as they had "the best facilities"
  - Luckily agreement was made for an independent neutral location – Telehouse
  - Telehouse was a Financial disaster-recovery centre – they took some serious persuading that this Internet thing was worth selling some rack space to

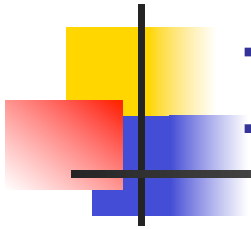
Success: UK peering

- LINX was established
  - Telehouse London
  - 5 UK network operators (4 commercial, 1 academic)
  - BTnet was a bit later to the party than the others
  - First "fabric" was a redundant PIPEX 5-port ethernet hub!
    - We had just deployed our first Catalyst 1201 in our PoPs
  - Soon replaced with a Catalyst 1201 8-port 10Mbps ethernet switch when the aggregate traffic got over about 3Mbps
  - Joined by a second one when redundancy and more capacity was needed



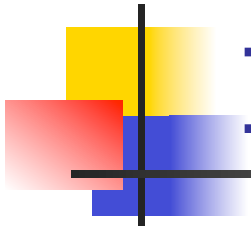
Third Lesson

- Peering is vital to the success of the Internet
- PIPEX sales took off
  - Customer complaints about RTT and QoS disappeared
  - Our traffic across LINX was comparable to our US traffic
- The LINX was critical in creating the UK Internet economy
  - Microsoft European Datacentre was UK based (launched in 1995), connecting via PIPEX and BTnet to LINX
  - Our resellers became ISPs (peering at LINX, buying their own international transit)
  - More connections: smaller ISPs, international operators, content providers (eg BBC)



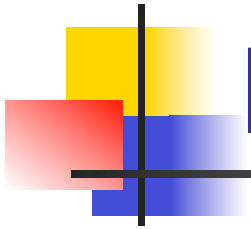
IGPs

- IGRP was Cisco's classful interior gateway protocol
- Migration to EIGRP (the classless version) happened many months after the Internet moved to BGPv4
  - Backbone point to point links were all /26s, and only visible inside the backbone, so the classfulness didn't matter
- EIGRP was Cisco proprietary, and with the increasing availability of other router platforms for access and aggregation services, decision taken to migrate to OSPF
  - Migration in itself was easy: EIGRP distance was 90, OSPF distance was 110, so deployment of OSPF could be done "at leisure"



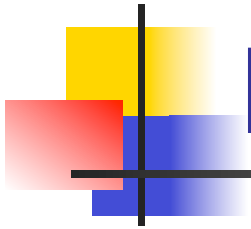
IGP migration

- IGP migration is generally simple, given each IGP has a different protocol distance
  - A path known via both EIGRP and OSPF sees EIGRP being preferred
  - When both protocols are operating, increasing EIGRP's protocol distance higher than OSPF ensures that OSPF takes over
  - Removing the old protocol is NOT such a good idea until:
    - All internal prefixes are in the new protocol
    - All connectivity is verified
    - The network has been operating as such for a period of time



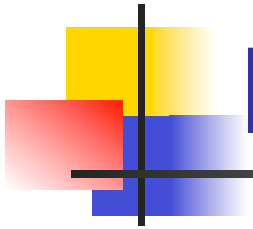
Fourth Lesson

- IGP migration needs to be done for a reason
  - With a documented migration and back out plan
  - With caution
- The reasons need to be valid
  - EIGRP to OSPF in the mid 90s took us from working scalable IGP to IOS bug central ☹ – the OSPF rewrite was still half a decade away
  - UUNET was by then our parent, with a strong ISIS heritage and recommendation
    - Cisco made sure ISIS worked, as UUNET and Sprint needed it to do so



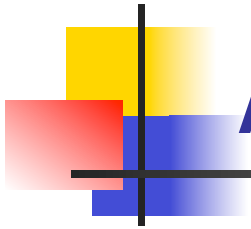
Redundancy

- A single link of course means a single point of failure – no redundancy
- PIPEX had two links from UK to US
  - Cambridge to Washington
  - London to New York
- On separate undersea cables
  - Or so BT and C&W told us
- And therein is a long story about guarantees, maintenance, undersea volcanoes, cable breaks, and so on



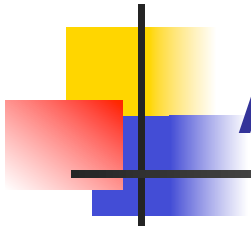
Fifth Lesson

- Make sure that critical international fibre paths:
  - Are fully redundant
  - Do not cross or touch anywhere end-to-end
  - Go on the major cable systems the supplier claims they go on
  - Are restored after maintenance
  - Have suitable geographical diversity (running in the same duct is not diversity)



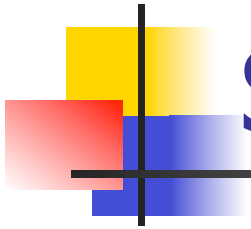
Aggregate origination

- Aggregate needs to be generated within ISP backbone for reachability
  - Leak subprefixes only for traffic engineering
  - "Within backbone" does not mean overseas PoP or at the peering edge of the network
- Remember those transatlantic cables
  - Which were redundant, going to different cities, different PoPs, diverse paths,…
- Having the Washington border routers originate our aggregates wasn't clever



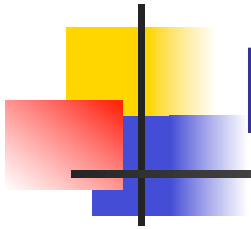
Aggregate origination

- Both transatlantic cables failed
  - Because one had been rerouted during maintenance – and not put back
  - So both our US circuits were on the same fibre – which broke
  - We didn't know this – we thought the Atlantic ocean had had a major event!
- Our backup worked – for outbound traffic
  - But nothing came back – the best path as far as the US Internet was concerned was via MAE-East and our UUNET peering to our US border routers
- Only quick solution – switch the routers off, as remote access wasn't possible either



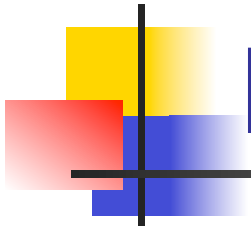
Sixth lesson

- Only originate aggregates in the core of the network
  - We did that, on most of the backbone core routers, to be super safe
  - **But never on the border routers!!**



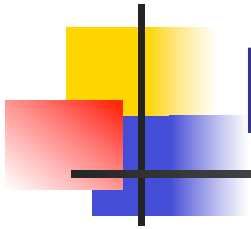
How reliable is redundant?

- Telehouse London was mentioned earlier
  - Following their very great reluctance to accept our PoP, and the LINX, other ISPs started setting up PoPs in their facility too
  - After 2-3 years, Telehouse housed most of the UK's ISP industry
- The building was impressive:
  - Fibre access at opposite corners
  - Blast proof windows and a moat
  - Several levels of access security
  - 3 weeks of independent diesel power, as well as external power from two different power station grids



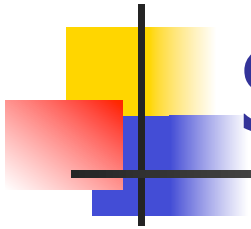
How reliable is redundant?

- Technically perfect, but humans had to run it
- One day: Maintenance of the diesel generators
  - Switch them out of the protect circuit (don't want a power cut to cause them to start when they were being serviced)
  - Maintenance completed – they are switched back into the protect circuit
    - Only the operator switched off the external mains instead
    - Didn't realise the mistake until the UPSes had run out of power
    - Switched external power back on – the resulting power surge overloaded UPSes and power supplies of many network devices
- News headlines: UK Internet "switched off" by maintenance error at Telehouse



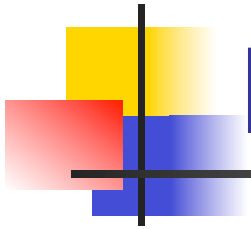
How reliable is redundant?

- It didn't affect us too badly:
  - Once BT and Mercury/C&W infrastructure returned we got our customer and external links back
  - We were fortunate that our bigger routers had dual supplies, one connected to UPS, the other to unprotected mains
    - So even though the in-room UPS had failed, when the external mains power came back, our routers came back – and survived the power surge
- Other ISPs were not so lucky
  - And we had to restrain our sales folks from being too smug
  - But our MD did interview on television to point out the merits of solid and redundant network design



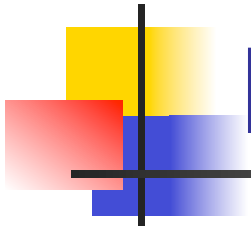
Seventh lesson

- Never believe that a totally redundant infrastructure is that
  - Assume that each component in a network will fail, no matter how perfect or reliable it is claimed to be
  - **Two of everything!**



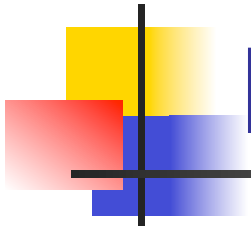
Bandwidth hijack

- While we are talking about Telehouse
  - And LINX…
- Early LINX membership rules were very restrictive
  - Had to pay £10k membership fee
  - Had to have own (proven) capacity to the US
  - Was designed to keep smaller ISPs and resellers out of the LINX – ahem!
  - Rules eventually removed once the regulator started asking questions – just as well!
- But ISPs still joined, many of them our former resellers, as well as some startups



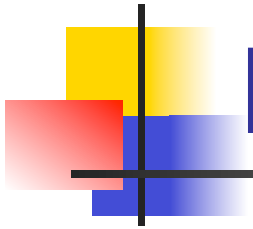
Bandwidth hijack

- We got a bit suspicious when one new ISP claimed they had T3 capacity to the US a few days after we had launched our brand new T3
- Cisco Netflow quickly became our friend
  - Had just been deployed on our border routers at LINX and in the US
    - Playing with early beta software again on critical infrastructure ☺
  - Stats showed outbound traffic from an AS we peered with at LINX was transiting our network to the US
  - Stats showed that traffic from an AS we didn't peer with at MAE-East was transiting our network to this same LINX peer
  - What was going on??



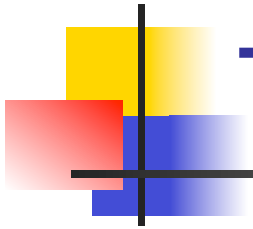
Bandwidth hijack

- What happened?
  - LINX border routers were carrying the full BGP table
  - The small ISP had pointed default route to our LINX router
  - They had another router in the US, at MAE-East, in their US AS – and noticed that our MAE-East peering router also had transit from UUNET
  - So pointed a default route to us across MAE-East
- The simple fix?
  - Remove the full BGP table and default routes from our LINX peering routers
  - Not announcing prefixes learned from peers to our border routers



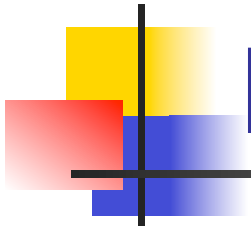
Eighth lesson

- Peering routers are for peering
  - And should only carry the routes you wish peers to see and be able to use
- Border routers are for transit
  - And should only carry routes you wish your transit providers to be able to use



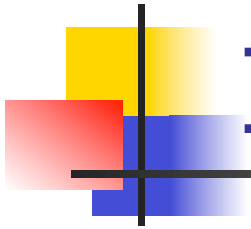
The short sharp shock

- It may have only been 5 years from 1993 to 1997
- But the Internet adoption grew at a phenomenal rate in those few years
- In the early 90s it was best effort, and end users were still very attached to private leased lines, X.25, etc
- By the late 90s the Internet had became big business
- Exponential growth in learning and experiences
  - There were more than 8 lessons!
- (Of course, this was limited to North America and Western Europe)



Moving onwards

- With UUNET's global business assuming control of and providing technical direction to all regional and country subsidiaries, it was time to move on
- In 1998, next stop Cisco:
  - The opportunity to "provide clue" internally on how ISPs design, build and operate their networks
  - Provide guidance on the key ingredients they need for their infrastructure, and IOS software features
  - All done within the company's Consulting Engineering function
- The role very quickly became one of infrastructure development

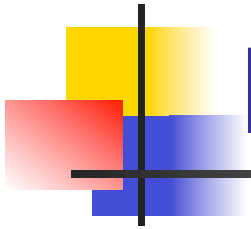


Internet development

- Even though it was only over 5 years, I had accumulated in-depth skillset in most aspects of ISP design, set up, and operational best practices
  - The 90s were the formative years of the Internet and the technologies underlying it
  - Best practices gained from experiences then form the basis for what we have today
- Account teams and Cisco country operations very quickly involved me in educating Cisco customers, new and current
- Working with a colleague, the Cisco ISP/IXP Workshops were born

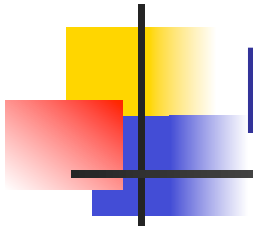
Internet development

- Workshops:
  - Teaching IGP and BGP design and best practices, as well as new features
  - Covered ISP network design
  - Introduced the IXP concept, and encouraged the formation of IXes
  - Introduced latest infrastructure security BCPs
  - Early introduction to IPv6
- Out of the workshops grew requests for infrastructure development support from all around the world



Development opportunities

- Bringing the Internet to Bhutan
- Joining AfNOG instructor team to teach BGP and scalable network design
- Introducing IXPs to several countries around Asia
- Improving the design, operation and scalability of service provider networks all over Asia, Africa, Middle East and the Pacific
- Helping establishing network operations groups (NOGs) – SANOG, PacNOG, MENOG etc
- Growing APRICOT as the Asia Pacific region's premier Internet Operations Summit



Bhutan

- In 1998, the 4th King decided that the Internet should be available in the country for the 25th anniversary of his coronation (2nd June 1999)
  - Technical staff from Druknet came to an ISP/IXP Workshop I ran with the UNDP in Malaysia in 1998
  - In March 1999 I received the call from UNDP in Bhutan asking for to help provide training for the Government's ISP
  - (And who would refuse, given Bhutan's status as one of the most reclusive and undeveloped countries in the world then)
  - There followed frantic activity in April before my trip there in early May

Reply

From Henrik Holde <henrik.holde@undp.org>
Subject **Bhutan: ISP setup**
To Philip Smith <pfs@cisco.com>
User agent Mozilla 4.04 [en] (Win95; I)

Philip - it was nice meeting you (although briefly) again at APRICOT in Singapore.

As you may be aware, Bhutan is about to make its way to the Internet and the first ISP in Bhutan will be funded jointly by UNDP Bhutan and APDIP. I am currently trying to identify means of providing training in various aspects of ISP management, routing, local access etc.

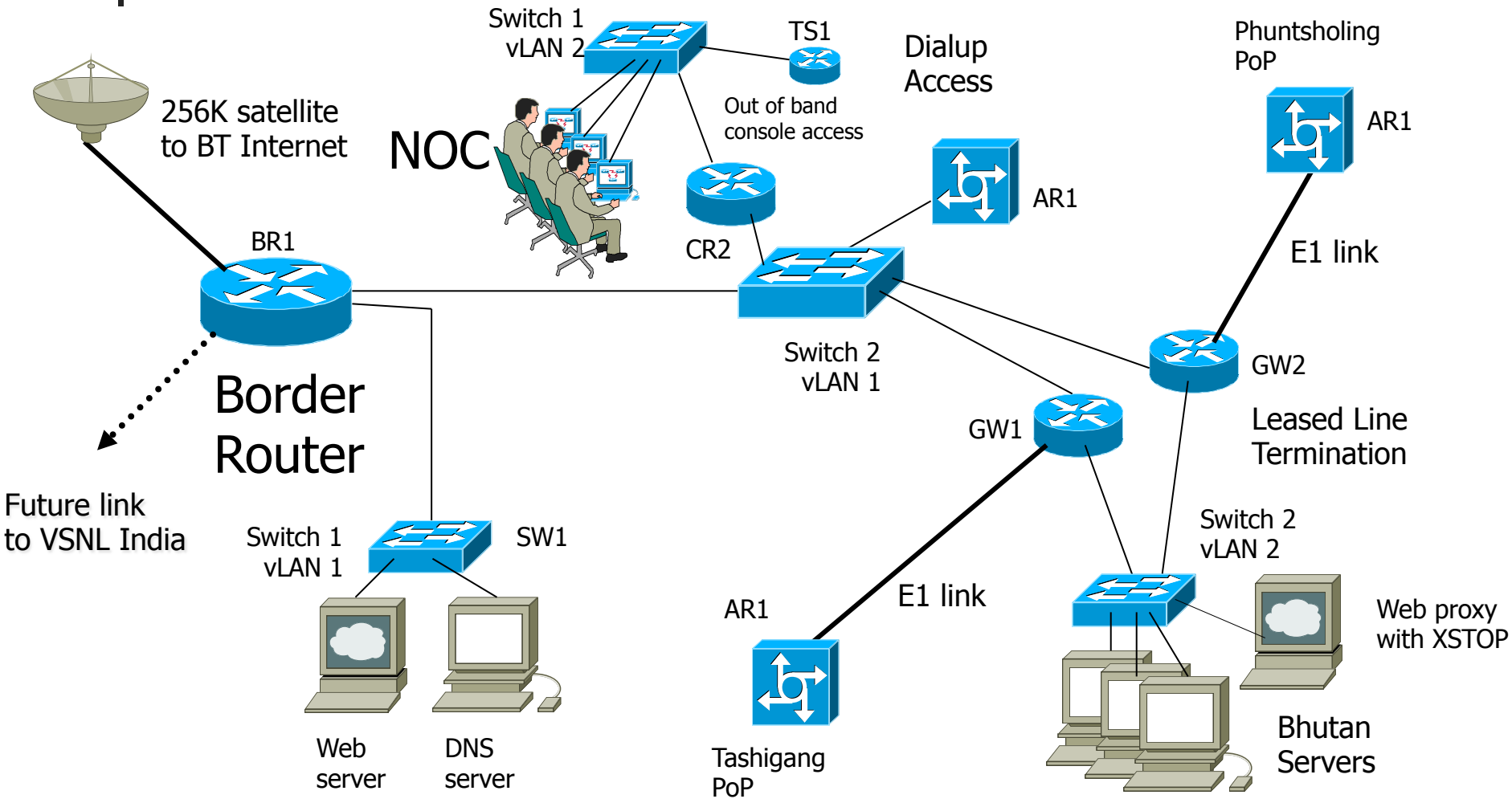
When we first spoke in November in Kuala Lumpur, you mentioned that you would be interested in visiting Bhutan. I was wondering if you would be able to combine a visit with providing some hands-on training in configuring and setting up the routing and local access equipment together with the Telecom/ISP staff here. We would obviously pay for your travel etc - unless Cisco would be interested in sponsoring your visit :-)~

I look forward to hearing from you - if you are interested in the above, the most likely time for the installation of the internet would be sometimes late April/early May.

Best regards,
Henrik

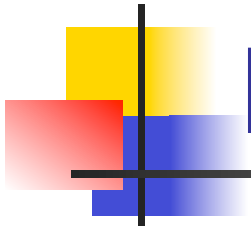
PS. There are literally no airconditioners in Bhutan!!

Network Diagram



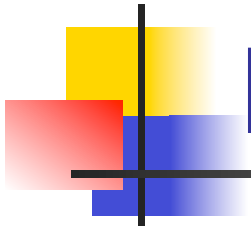
Bhutan in 1999

- Network looks a bit messy in retrospect:
  - But this was a rescue job
  - Used whatever equipment had already been delivered
    - (Cisco 2511 access servers, IBM AIX Servers)
  - Plus Cisco routers/switches specially purchased for this job
  - No time for refinements!
- Designed and built as an ISP
  - 256kbps satellite link to UK
  - Dialup via Cisco 2511 and modems
  - Leased line access via Cisco 3640
  - Border router was Cisco 2611
  - Replaced previous "Internet Café" design proposal



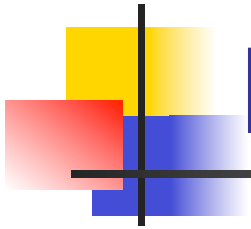
Bhutan in 2013

- International fibre:
  - 2.5Gbps to London
  - 2.5Gbps to Hong Kong
  - 1 Gbps to Chennai
- National IPv6/IPv4 backbone
- Redundant fibre and radio links
- Redundant and scalable PoP architecture
- Wide roll out of broadband and mobile data access
- Coverage in most districts (even though many don’t have road access)
- 3 other competing ISPs
  - Still no IXP – sigh!



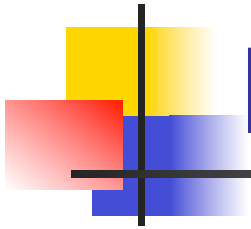
Nepal's IXP

- In 2002 Nepal had no IXP:
  - Nepal Telecom providing internet access
  - A few ISPs with their own satellite links
  - Mercantile & Worldlink providing transit to some smaller ISPs
  - No domestic traffic; traffic between ISPs went via Europe or Hong Kong
- Following the inaugural SANOG in Kathmandu, NPIX was launched, with agreement from some of the ISPs
- A tall building was found (the location had one small ISP – EverestNet)
  - Tall -> wireless would be the primary means of access



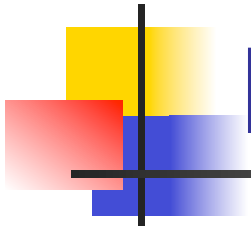
Nepal's IX

- In the months after SANOG 1, NP-IX was launched, established, switch installed, and the initial connections made
- Nepal Telecom refused to participate as they were the Govt and National Carrier
  - The independent ISPs carried on regardless
- Most problems were about getting the other ISPs connected
  - Wireless interference, line of sight, etc
- Configurations:
  - Even though a BGP/IXP Workshop had been run, routing knowledge was limited



Nepal's IX – configurations

- Getting the IXP running took persistence!
- Spent a week with Gaurab Raj Upadhaya driving around Kathmandu, visiting ISPs:
  - Much time spent sitting in traffic jams
  - Procuring ASNs from APNIC
  - Deploying BGP (iBGP, eBGP)
  - Fixing broken routing
  - Replacing static routes with OSPF
  - Upgrading router software
  - Giving impromptu crash courses in BGP and OSPF
  - etc



NPIX today

- Nepal Telecom finally agreed to join
  - Pressure from their customers as most local content repatriated, and now hosted on ISPs connected to the IX
- IXP now in two locations in Kathmandu
  - Considered vitally important national infrastructure
  - Traffic peaks at 300Mbps
  - www.npix.net.np

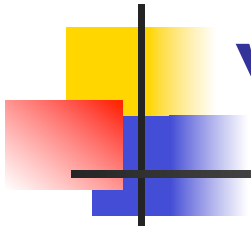
IXPs in general

- Establishing IXPs in a country always has its own set of stories
- Sadly many countries around the world are without any Internet Exchange Point
  - Some are too small, having only one or two viable ISPs
  - Others are bigger, and the quality of Internet and of Internet access is very low
  - IXPs need to come from a desire within the industry – outside folks can only explain the stunning benefits
  - If Vanuatu (small Pacific island nation) can justify an IXP, and see the benefits, almost every other country can too

Mongolia

- Long association with Mongolian industry, from that same UNDP workshop in 1998
- First workshop on-site in Ulaanbaatar after ISPs experienced problems with "the Internet disappearing"

- Shipping workshop equipment was one story!
  - Flights and aircraft hold sizes do matter – workshop kit box was 3cm too tall to fit into a Boeing 737, so the weekly Korean Airlines Airbus 300 had to be it
- Doing the workshop (with Gaurab) was something else

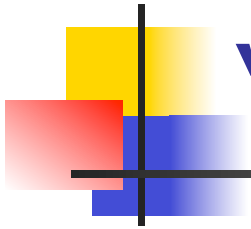


“The disappearing Internet”

- What was that about?
- BGP was set up for the main ISP in 2000 by an engineer flown in by Cisco
  - It was very well done, but…
  - The ISP was experiencing problems, with customer complaints, couldn’t access CNN, BBC, and some other major international media websites
- Geography: Mongolia is sandwiched between Russia and China
  - Transit only available via those two countries, or by satellite

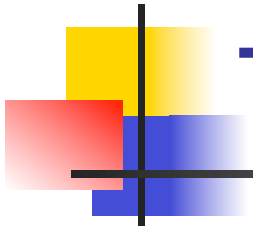
“The disappearing Internet”

- The only way in or out is through China or Russia
- Suspicion lay with the “Great Firewall of China”
  - The ISP got BGP transit from a Chinese ISP
  - Even though their upstream denied this
  - Not much love lost between the two countries



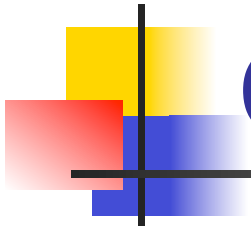
"The disappearing Internet"

- The GFW reason seemed somewhat unlikely – plausible, but unlikely
- What was happening between 2000 and 2005?
  - Significant growth of content distribution networks
  - Significant growth in distribution of new address space
  - Combining the two: new content networks were using new address space
- The disappearing Internet were the BGP filters put in place in 2000:
  - All IANA unallocated address space had been blocked in those filters
  - **Removing the filters (BGP and static null routes) made the Internet "reappear" again!**



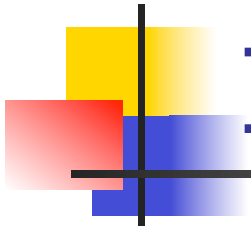
The lesson

- No matter how fantastic a reason for failure might seem, the real reason will be more mundane
- The real lesson:
  - Don’t use static filters to block unused address space without keeping it up to date
  - Folks like Team Cymru offer a BGP feed – much easier for maintenance!
- The other lesson:
  - Learn BGP for yourself rather than outsourcing – it’s not that hard



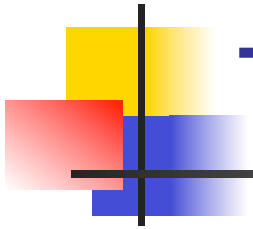
Ghana

- This goes back to 1993 – my first international customer at PIPEX
  - NCS had a Sun workstation (a 4/110 ?), running MorningStar PPP
  - ftp://ftp.funet.fi/pub/netinfo/dialup-ip/MorningStar/ppp.old/user-guide.ps.Z
  - Fixed analogue line from Accra to Cambridge – 2400bps!
  - Keeping that link going was almost a full time job
    - Power outages in Ghana
    - Inexplicable outages on the analogue link
    - Many phone conversations with William Tevie and Nii Quaynor
    - Interoperability between Telebit Netblazer PPP and Morningstar PPP kept me busy with both companies!
  - NCS's Sun (austin.gh.com) ran the DNS for .gh, as well as email for all of Ghana



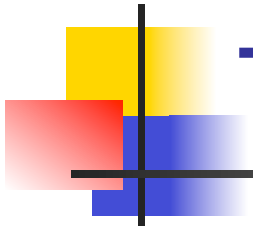
Is this the final lesson?

- Having two vendors involved means open season in finger pointing
  - PPP was RFC1331 in 1992, updated December 1993 (RFC1548) and then in July 1994 (RFC1661)
  - Many excuses for lack of interoperability
- Dual vendor strategy can be useful to avoid dependencies
  - Make sure both vendors know that they are responsible for problem resolution, and that you are not the referee



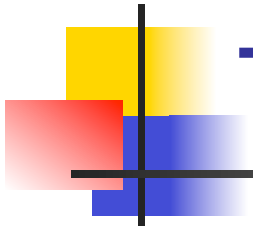
The story goes on…

- Other IXP experiences
  - Bangladesh, Singapore, Vanuatu, India, Pakistan, Uganda, PNG, Fiji, Samoa, Thailand, Philippines,…



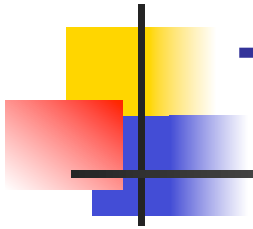
The story goes on…

- Other ISP design and redesigns



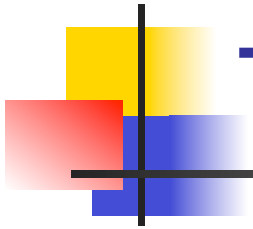
The story goes on…

- Satellites
  - falling out of sky
  - latency/tcp window vs performance



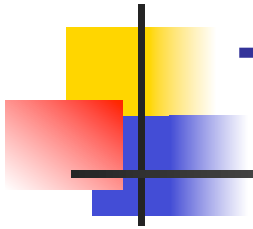
The story goes on…

- Fibre optics being stolen
  - Folks thinking it is copper



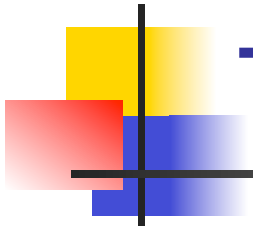
The story goes on…

- The North Sea fogs and snow which block microwave transmission



The story goes on…

- "You don't understand, Philip"
  - From ISPs, regulators, business leaders, who think their environment is unique in the world



The story goes on…

- "Ye cannae change the laws o' physics!"
  - To operators and end users who complain about RTTs

§ Montgomery "Scotty" Scott: Star Trek