Internet Development Experiences and Lessons

Philip Smith

BDNOG 1

23rd May 2014

Dhaka

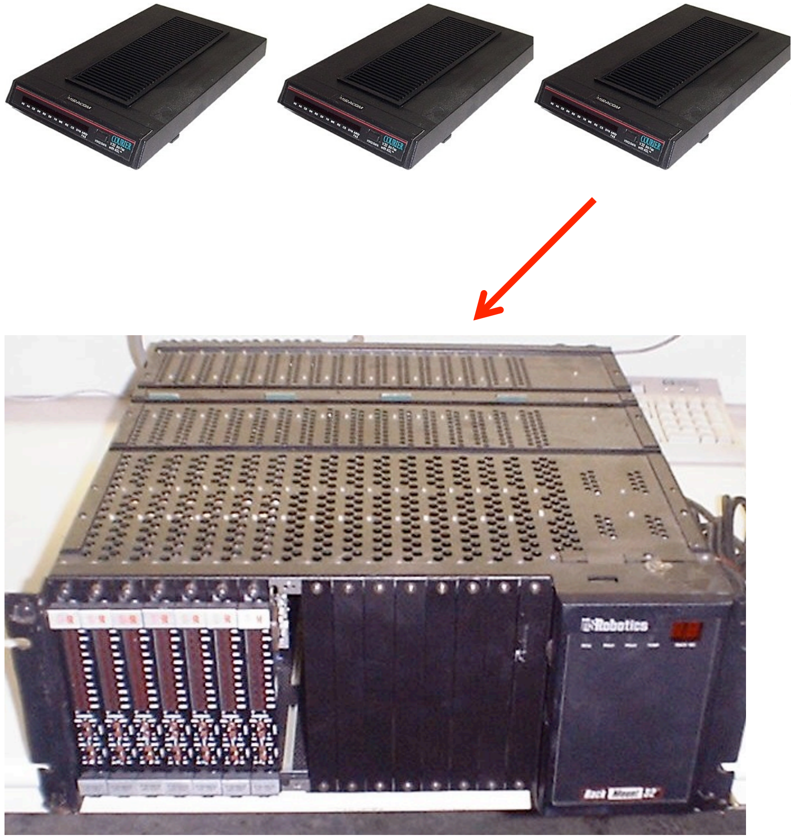
Background

- 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 2014 has never had it so good!
- In 1994/5:
  - iBGP was fully meshed
  - Routers had 16Mbytes 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"

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 ☹ – Cisco's 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

Network 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’s 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 a customer of ours also present at LINX (we didn’t peer with customers) was transiting our network via LINX to the US
  - Stats showed that traffic from an AS we didn’t peer with at MAE-East was transiting our network to this customer
  - 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 ISP 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

NOG Development

- Started getting more involved in helping with gatherings of local and regional operations community
  - APRICOT was the first experience – difficulties of APRICOT '98 and '99 led to a refresh of the leadership in time for APRICOT 2001
  - APRICOT growing from strength to strength – but annual conference had 56 economies across AsiaPac to visit!
  - Regional and Local NOGs were the only way to scale

NOG Development

- NZNOG and JANOG were starting
- SANOG launched in January 2003, hosted alongside Nepalese IT event
  - Several international "NOG experts" participated
  - Purpose (from www.sanog.org):

> SANOG was started to bring together operators for educational as well as co-operation. SANOG provides a regional forum to discuss operational issues and technologies of interest to data operators in the South Asian Region.

    - And this is a common theme for most NOGs founded since

Ingredients for a successful NOG

① Reach out to community and organise a meeting of interested participants

② Reach out to colleagues in community and further afield and ask them to come talk about interesting operational things

③ Figure out seed funding and find a place to meet

④ Commit to a 2nd NOG meeting

⑤ Have fun!

Ingredients for a successful NOG

- Avoid:
  - Setting up lots of committees before the NOG meets for the first time
  - Worrying about what fees to charge or discounts to provide
  - Worrying about making a profit
  - Hiring expensive venues, event organisers
  - Providing expensive giveaways
  - Providing speaking opportunities to product marketeers

Ingredients for a successful NOG

- During that first meeting:
  - Solicit suggestions about the next meeting
    - Location, content, activities
  - Suggest a mailing list
    - And then set it up, encouraging participants to join
  - Encourage organisations participating to consider future sponsorship
  - Encourage colleagues to help with various tasks
  - Organise a meeting of the folks who helped pull the first meeting together
    - Here is the first committee, the Coordination Team

Ingredients for a successful NOG

- After the first meeting:
  - Plan that 2nd meeting, relaxation is not allowed
  - Don’t expect lots of people to rush and help
    - NOG leadership is about being decisive and assertive
    - And can often be lonely
  - Organise the next meeting of the Coordination Team (face to face, teleconference,…)
  - Don’t lose momentum
  - Keep the Coordination Team involved

Ingredients for a successful NOG

- Going forwards:
  - Encourage discussion and Q&A on the mailing list
    - No question is too silly
  - Run the second meeting, plan the third
    - But don’t try and do too many per year – one or two are usually enough
    - Don’t rely on the international community for everything – encourage and prioritise local participation
    - Start thinking about breaking even
  - After the 2nd or 3rd meeting, assistance with programme development – the next committee!

The final lesson?

- Setting up a NOG takes effort and persistence
  - Bring the community along with you
  - People attend, and return, if the experience is positive and the content is worth coming for
  - Include all sectors and regions the NOG claims to cover
  - Budget needs to be neutral, sponsorship generous, participant costs low
  - No bureaucracy!

The story goes on…

- IXP experiences
  - Nepal, Bangladesh, Singapore, Vanuatu, India, Pakistan, Uganda, PNG, Fiji, Samoa, Thailand, Mongolia, 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