Analysis of Design Decisions in a 10G Backbone

Vijay Gill

<vijaygill9@aol.com>

Why

- Vegas NANOG-reform meeting
  - How to build a backbone in 3 Slides

Slide 1

Hire RFP Engineers

“Oi! Vendor, run me up a backbone here, then”

Heidi Heiden's First Law:
When you want it bad, you get it bad,
and most people want it in the worst way.

Slide 2

Slide 3

PROFIT

The Good Old Days

- FRITCH (Frame Relay Interim Crutch)
- 3com cards with 2KB of packet memory.
- IOS upgrades: Call Cisco. Start upgrade.
- Routing updates stopped packet forwarding
- OC12s built on protect side
- Two words: DEC Gigaswitch
- LS2010s that took 40 minutes to boot
- FORE ASX memory leak
- TTM linecards

Pittsburgh construction contractors drop packets more than a cisco 2501 which is running “we poured coffee in the aui port” instead of IOS

-Faisal Jawdat

Today

- No more fundamentally broken stuff
- Routers are not running out of PPS, obviating need for things like full-mesh ATM
- Methodology is now understood

'Everybody wants to be a bodybuilder, but don't nobody wanna lift no heavy ass weight!' - Ronnie Coleman

AOL

Guiding Principles - Efficiency

- Companies are running based on models that are predicated on 50% or more revenue models
- Those days are over
- The killer App is bandwidth
- Become the Walmart ™of the internet
- Learn to survive based on 10% profit margins. Or lower.
- OSS/NMS are competitive advantages

“it's the bandwidth, stupid”
-John McCalpin

Guiding Principles - Real Options

- Theory of Real Options
  - Market uncertainty is high
  - Architecture that fosters experimentation at the edge creates potential for greater value than centralized administration
  - Distributed structure promotes innovation
  - Enables experimentation at low cost
- Putting the intelligence in the applications
- Spend capital on improving backbone or making application more resilient?
  - Pick application
    - Spend can be leveraged across the entire Internet
    - Can remove backbone

AOL

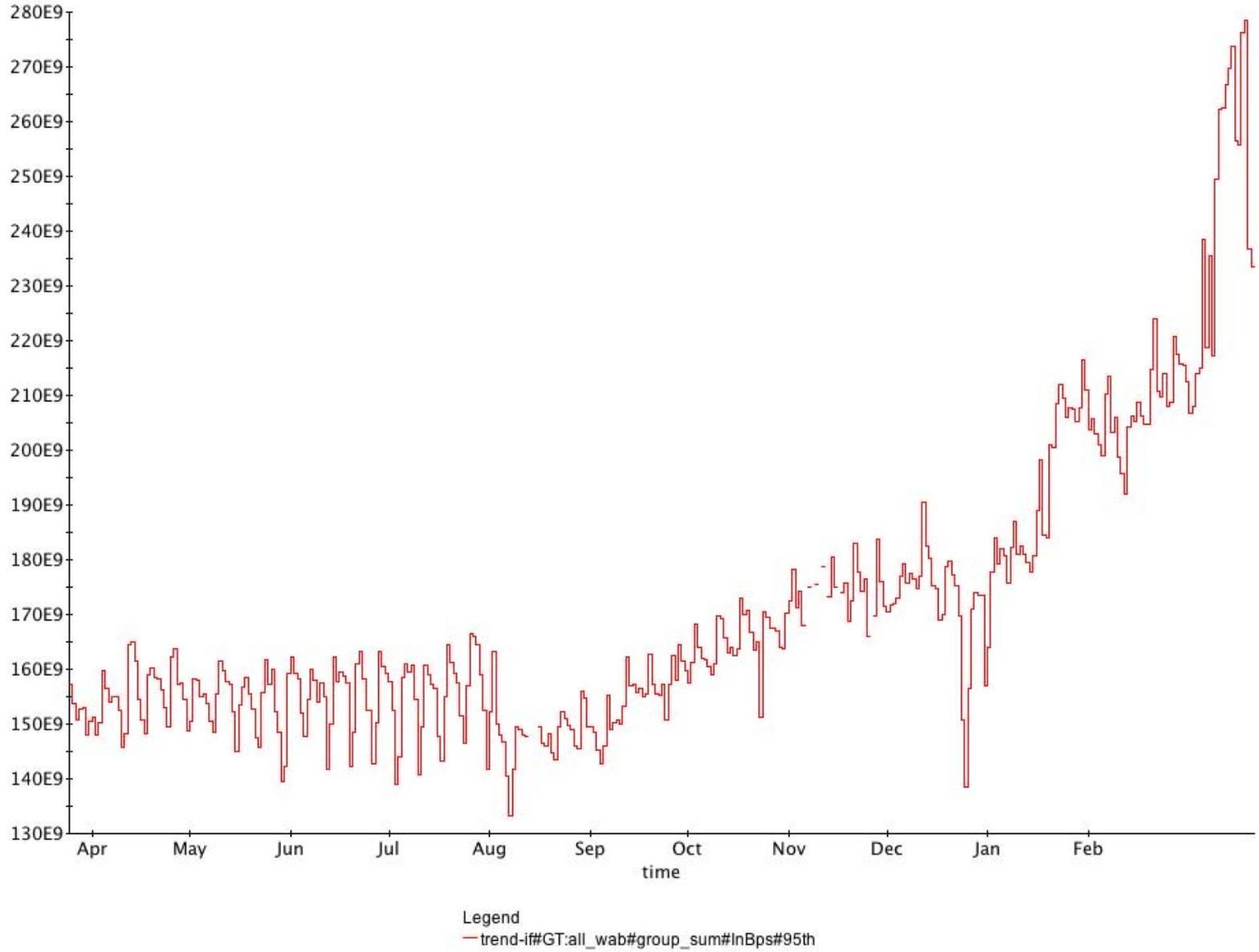
What Is ATDN?

- International IP Network that connects all AOL/TW facilities to the internet
  - OC192 Backbone in US and Europe
- Provides:
  - AOL/TW Facility Interconnections
  - Internet Connectivity
  - AOL Broadband/Dial Access Connectivity
  - Inter-Data Center Connectivity

So I've been thinking of starting a company. I need to write a business/marketing plan. Also, I need crack, a BMW 540i, and chix0rs

-Joshua Schachter

Traffic Volume



What Are The Bits?

- 50 Gigabit/sec of edge traffic
- 21+ million AOL subscribers
- 3+ million CompuServe subscribers
- 3+ million TW Cable subscribers
- 125+ million AIM users worldwide
- 110+ million ICQ registered users
- Peak simultaneous usage:
  - 3.1M AOL users online
  - 7.3M active AIM users
  - 2.4M active ICQ users

Infrastructure:

- 25,000+ host servers
- ~ 500,000 sq ft raised floor space
- ~ 800 optical backbone routers
  - ATDN – 66+ 10-gig capable routers in the backbone
  - ATDN – 100+ edge routers
- ~ 3000 L2/L4-7 Switches
- 66,000 interfaces polled every 5 mins
- Over 1 million network variables captured

AOL

Accounting Policy

- OIBDA
  - Cash related expenditures
    - Personnel expenses (SG&A)
    - Monthly facility costs – Rent, Power, Remote Hands etc.,
- Non-OIBDA
  - Depreciable assets – Equipment, IRU's (if they meet certain policy thresholds)
- The Line is Expense vs. Depreciation
- Below the line
  - Depreciation/Amortization. Non-cash. Capital. Equipment that is being depreciated.
- Above the line
  - SG&A. Cash based transactions, MRCs, expenses, rents
- Operating Lease: Cash transaction.
- Capital Lease: Amortizable, 15 years or more depending.

It is hard to sell ATM services when your switches are locked in NON-CONFORM cages across the world.

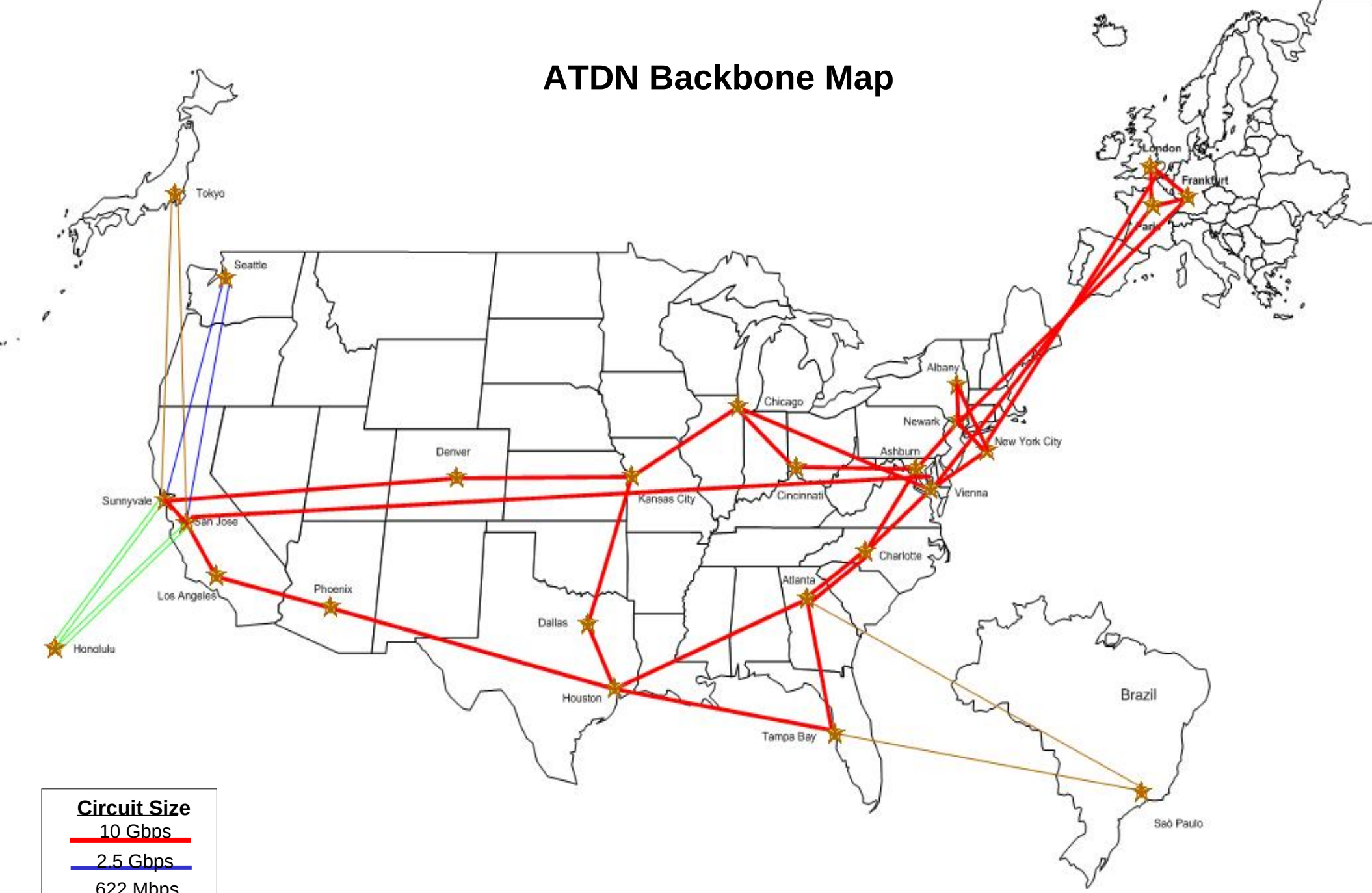
-Adam Rothschild

Build Vs. Buy

- Comparison of purchased lit capacity (monthly lease) versus organic build out
- Organic build out scenario includes:
- Procurement of fiber and equipment – Sunk cost
- Ongoing monthly expenses - Collocation, Equipment Maintenance, SG&A
- NPV – To determine feasibility of the project
- Factors include:
  - Weighted Cost of Capital – Borrowing rate of Capital
  - Cash Flow
- Payback within 18/24 months
- Best financial decision

It was, in Wall Street terms, fully funded.
-Jim Jubak

ATDN Backbone Map



Circuit Size

- 10 Gbps
- 2.5 Gbps
- 622 Mbps
- 155 Mbps

ATDN Design Philosophy

Complete life cycle in NetOps

- Architect, design, implement, and operate

Design Criteria

- Diversity of component, paths, logical units
- No Single Point of Failure (SPOF)
  - Redundant capacity to support peak load
- Routed topology
- System review to ensure performance & survivability goals

"Fault tolerant" is like "tier 1" -- the companies that really are fault-tolerant aren't the ones going all over the place claiming to be.

-Robert Seastrom

AOL

Design Goals – Operational Simplicity

- Standards based (as much as possible)
- Consistency
  - Replicate design throughout all POPs
    - Boxes might be different sizes but nodal architecture is the same
- “Choose the path ***more*** traveled”
  - Problem with being on the cutting edge is that someone’s got to bleed
- Empower new employees with the ability to easily understand what has been built
  - Push tasks into the NOC

“Do not be so proud of this technological terror you have constructed. The ability to criticize Star Wars is insignificant next to power of the Fans”

-Brandon David Short

AOL

Design Goals – Operational Simplicity

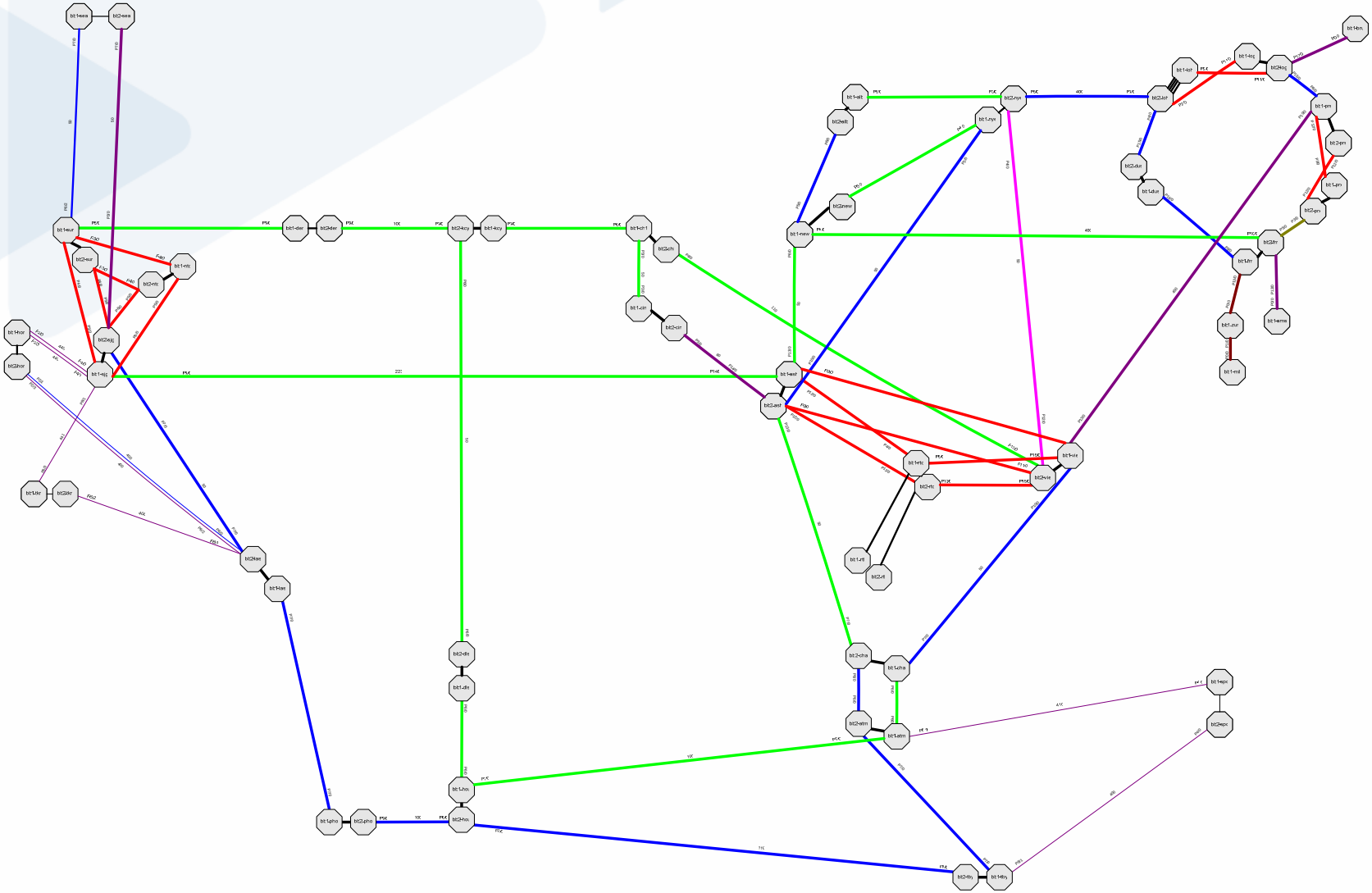
- When to touch the network
  - Routing policy based on simple performance metrics and cost (price/Mb)
  - No fancy tricks – Traffic engineering based on using common policy across similar peers (e.g., free or transit) rather than exception policy
- Achieved through engineering simplicity
  - Focus on reducing OPEX
  - It's not about building a network that's cool or is a challenge to engineer
    - "Make it simple, but no simpler"

Ask not what evolution can do for you, ask what you can do for evolution.

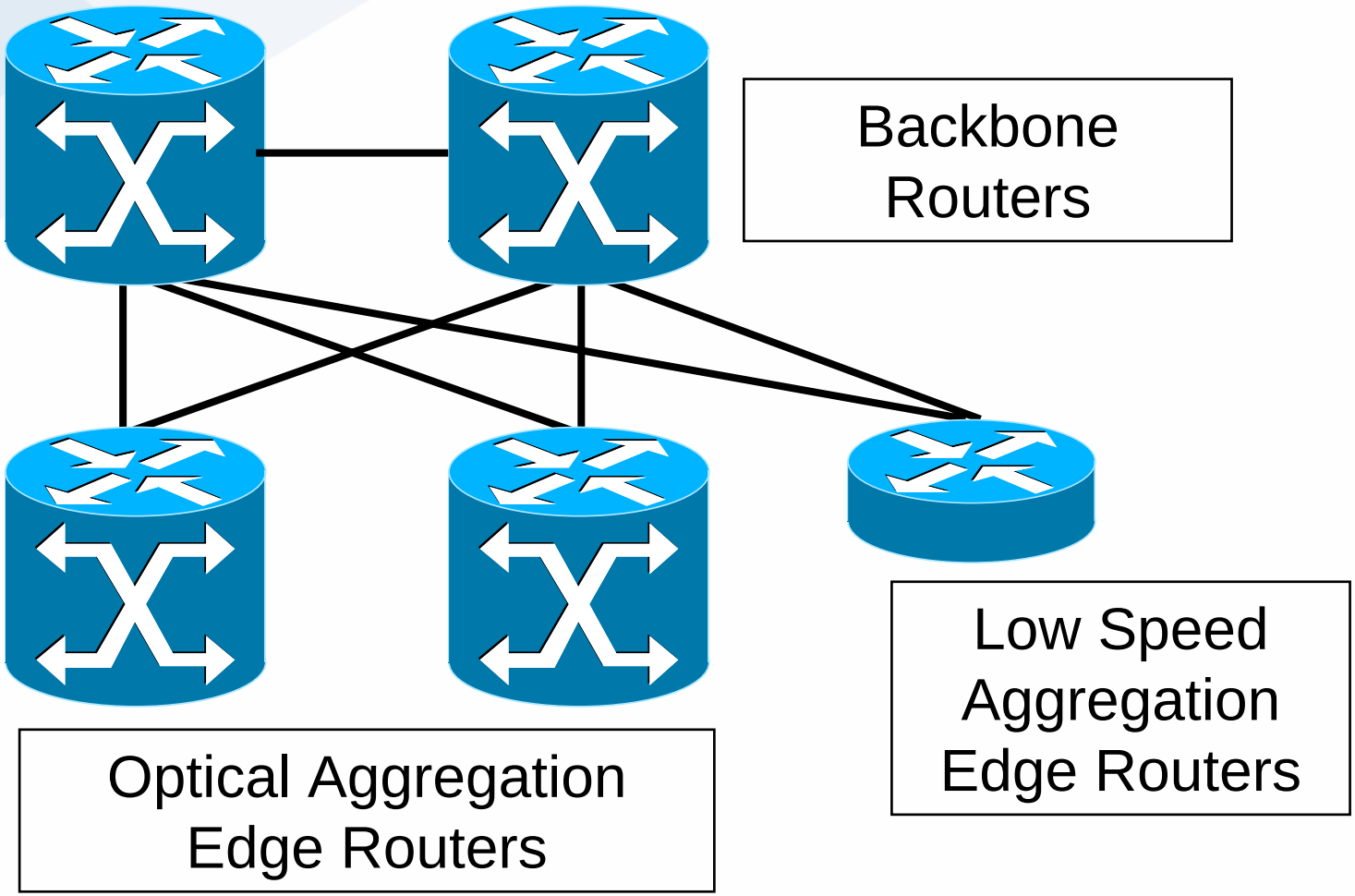
- Jimbo Kukla

AOL

ATDN Logical Design



ATDN Hub Design



ATDN Architecture

- Two Level Hierarchy
  - Edge and Core
- Optimized for minimal cost of operation (OPEX and SG&A)
- Overprovision bandwidth (for now)
- Scaleable

The internet doubles in size every x months. the amount of clue to run it remains the same.
-Bill Manning

Forced Hierarchy

- Separation of routing functions in the core and the edge
- Well-defined interfaces between the edge and core
  - Easy to isolate problem equipment, circuits, or traffic sources/sinks
  - Reduces risk and complexity of substituting substantially different hardware or software in the network as it evolves
- Network subsystems can be evolved separately as long as the interface presented remains unchanged

ATDN Design Highlights

- Core consists of large transit hubs
  - Yields good aggregation
  - Having fewer transit hubs reduces the number of transit paths in the network
    - Prevents combinatorial explosion of possible paths
    - Traffic engineering is relatively simple

"But in our enthusiasm, we could not resist a radical overhaul of the system, in which all of its major weaknesses have been exposed, analyzed, and replaced with new weaknesses."

-Bruce Leverett

AOL

Routing Architecture

- BGP/IS-IS
- Run IGP lean, carry everything in iBGP
- IGP cost structure:
  - Local
  - Regional
  - Long-haul
  - International

Sir,
Did you ever have the pleasure of implementing IS-IS's and OSPF's flooding algorithms? I did.
I can tell you from experience: OSPF is just a %$~@!-ed up protocol.

-Henk Smit

AOL

Routing Architecture Robustness

- No protected rings at the SONET level
- Take unprotected bandwidth
- Rely on IP for protection
- No single failure of any component should isolate a major hub
- Primary, secondary and tertiary backup paths

"When you're good and crazy, the sky's the limit!!"
-The Tick

Traffic Engineering

- Matrix of city-pair flows
  - Macro flows are tractable
  - Watch bandwidth usage
- Capacity Planning
  - Add capacity where and when required
  - Increase primary and backup capacity in sync
- Egress capacity is the bottleneck

“The superior pilot uses his superior judgement to avoid situations in which he has to demonstrate his superior skill”

-Unknown

AOL

Routing Policy

- Standard Filtering (external)
  - Nothing longer than /24, no RFC1918, no private ASNs, etc., no martians
  - Full edge packet filtering on AOL space as well as RFC1918 source/dest
  - All routes tagged with communities to record origination
  - Customers are filtered by prefix
- Peers
  - No prefix/as-path filters applied to peers other than above
  - Maximum prefix used as a coarse sanity-check

AOL

Control Plane Policing

```
ip icmp rate-limit unreachable 5
 ip cef linecard ipc memory 10000
 no ip source-route, finger etc
 ip receive access-list 4505
 !
```

```
access-list 4505 permit ip a.b.c.d 0.0.0.7 any
  access-list 4505 permit tcp any any eq bgp
  access-list 4505 permit tcp any eq bgp any
  access-list 2405 permit ospf any any
  etc…
```

This calls for a very special mixture of psychology and extreme violence.

- Vyvyan, The Young Ones

AOL

Importing Routes into BGP

Desired Goals

1. Minimal manual intervention
2. Consistency
3. Flexible

Note 1 and 3 appear contradictory

```
router bgp 1668
   no synchronization
   bgp log-neighbor-changes
   bgp deterministic-med
   bgp bestpath compare-routerid
   bgp dampening 10 2000 4000 30
   redistribute connected route-map FR-CONNECTED
   redistribute static route-map FR-STATIC
```

AOL

```
route-map FR-CONNECTED deny 10
   description deny all unwanted routes
   match ip address prefix-list FR-CONNECTED-1
  !
  route-map FR-CONNECTED permit 20
   description public specifics - not to the Internet
   match ip address prefix-list FR-CONNECTED-2
   set local-preference 300
   set origin igp
   set community 1668:2xxxx
  !
  route-map FR-CONNECTED permit 30
   description Catch all
   set local-preference 300
   set origin igp
   set community 1668:2yxxx
  !
```

How Did We Do?

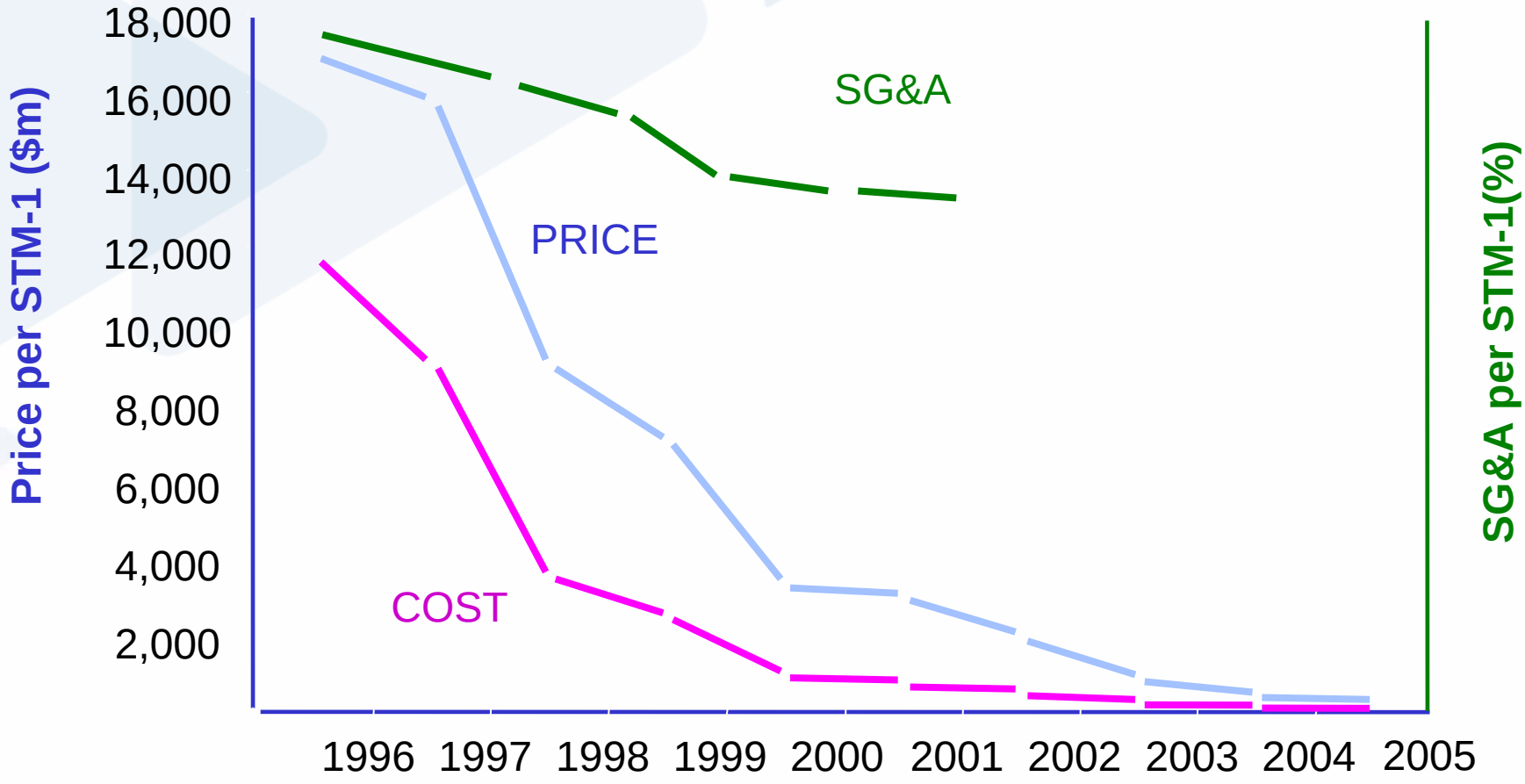
- Traffic volume has grown 230% (75 to 250 G/s)
- SG&A has gone down (staff reduced 34% over same period)

Chain gangs and slave labor can cut the costs considerably. And, when you no longer have a use for them, they are completely biodegradeable.

-John Jasen

AOL

Cost Per Bit – Major Components

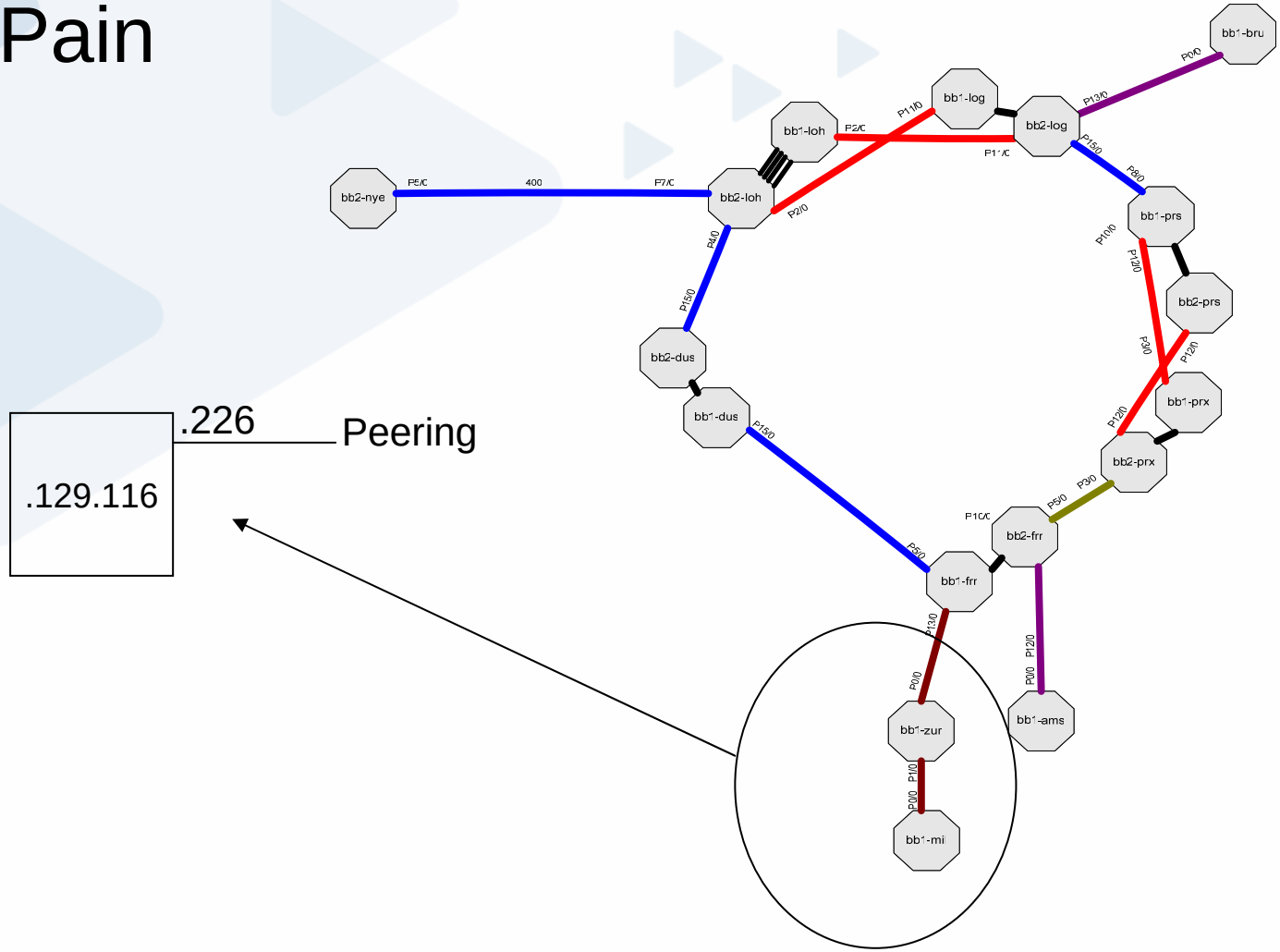


Historical and forecast market price and unit cost of Transatlantic STM-1 circuit (on 25 year IRU lease)

Source: KPCB

AOL

Pain



There is a difference between making something foolproof and reducing the number of fools"

-Bill Barns

```
pop1-ash#show ip route 66.185.138.226 [peering interface]

Routing entry for 66.185.138.224/30 [peering interface]

  Known via "isis", distance 115, metric 1478, type level-2
    Redistributing via isis
    Last update from 66.185.139.192 [bb1-ash]
    Routing Descriptor Blocks:
    * 66.185.139.192, from 66.185.129.116 [bb1-mil] , via POS0/0

pop1-ash#show ip route  66.185.129.116 [bb1-mil]
  Routing entry for 66.185.129.116/32
    Known via "isis", distance 115, metric 975, type level-2
    Redistributing via isis
    Last update from 66.185.139.192 [bb1-ash] on POS0/0, 13:42:55 ago
    Routing Descriptor Blocks:
    * 66.185.139.192, from 66.185.129.116
```

BGP

- Maintenance on routers in Milan
- Did not shut down .226 peering
- No reachability to next-hop of milan-bb
- Decided to follow nailed up route

"When your hammer is C++, everything begins to look like a thumb."

-Steve Hoflich

AOL

IGP reachability for bb1-mil goes away

iBGP has not timed out

Next-hop now resolves to nailed up static blocks for AOL space

This is local

- All NLRI coming from .226 are now reachable with a local next-hop

AOL: What happen?

BGP: I am in your base, killing all your dudes

Pagers: someone set us up the bomb

Mitigation

```
router isis
 net 39.752f.0100.0014.0000.5000.1668.0661.8512.8107.00
 is-type level-2-only
 domain-password you’ve-got-mail!
 metric-style wide
 external overload signalling
 set-overload-bit on-startup wait-for-bgp
 max-lsp-lifetime 65535
 lsp-refresh-interval 65000
 no hello padding
 log-adjacency-changes all
 redistribute static ip route-map FR-STATIC-LB
 passive-interface Loopback0
 maximum-paths 6
```

AOL

Mitigation

```
ip prefix-list FR-STATIC-LB permit 66.185.128.0/23

  route-map FR-STATIC-LB permit 10
   description Mitigate blackholing when edge next-hop dies
  without being withdrawn from IBGP - PR
   match ip address prefix-list FR-STATIC-LB
   set metric 500000
   set metric-type internal
   set level level-2
  !

  ip route 66.185.128.0 255.255.254.0 Null0
```

Q&A

- Questions

"The venture will be profitable from day one"
-Michael Armstrong (announcing the Concert deal)