Operational Feedback to IP Equipment Vendors

Vijay Gill

vijaygill9@aol.com

NANOG 26, Eugene, OR

October 26,2002

Audience

- Targeted at vendors of IP equipment used by ISPs
- Telecom Meltdown
- Went From
  - *Damnit. how many more core router startups can there be? F!@#$g half my email box with every type of tree, bush, shrub, and fruit.*

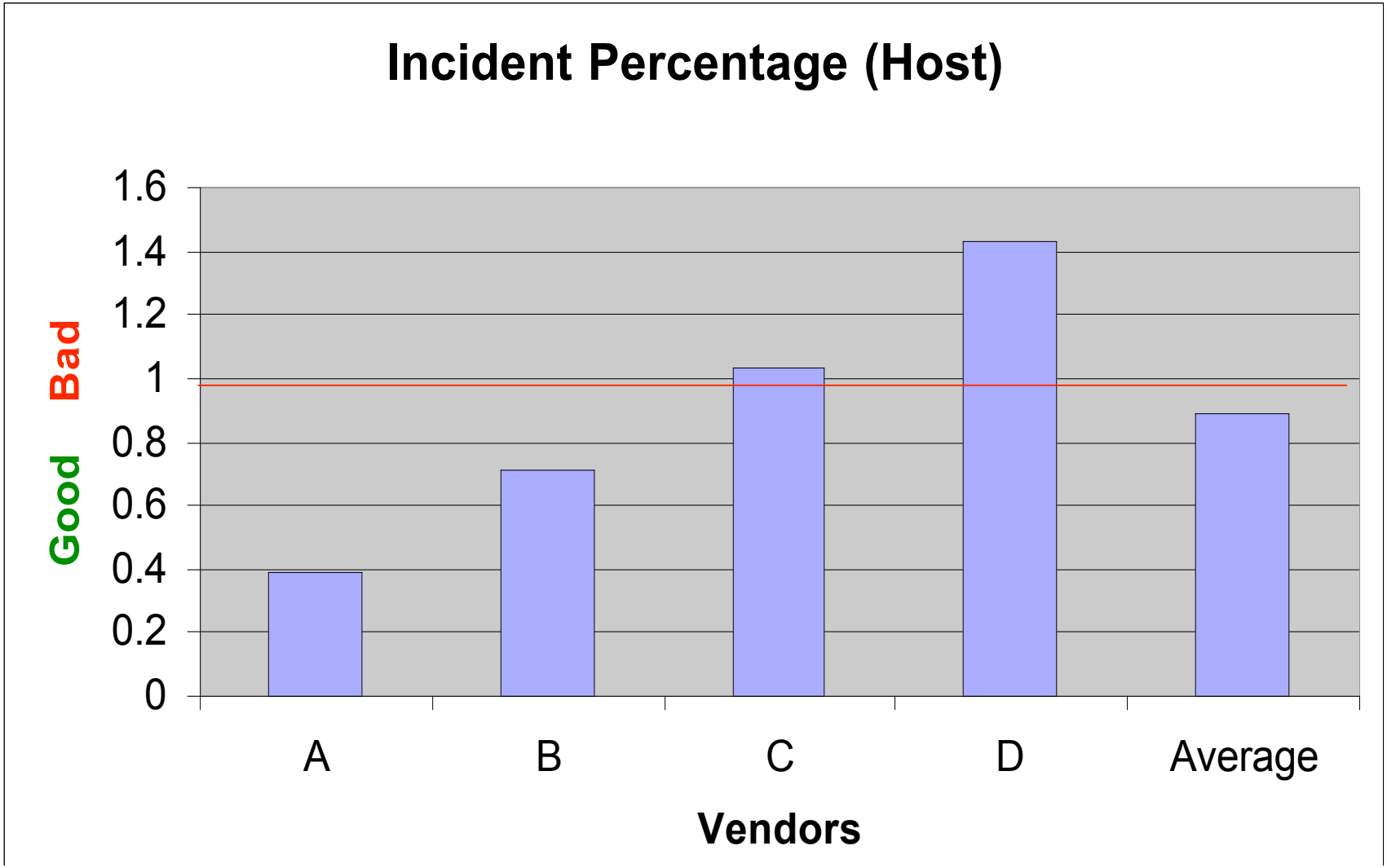
    -Dave Cooper (1999)
- To
  - *The unemployment office only gives money, not options.*

    -Bill Fumerola (2001)
- Targeted at ISPs
  - The CAPEX Hammer
- Where the problems are
  - Networks cost a lot to run
    - Need to focus on reducing Capital and OPEX
  - Security

Costs

- We can only squeeze costs so far
  - Bandwidth and hardware costs are more elastic
- Human cost remains constant
- Need more robust software and hardware
  - Excessive complexity isn't going to get us there
- Each incident costs money
- Chart on right shows some vendors are more expensive (OPEX)

Hosts (For Comparison)

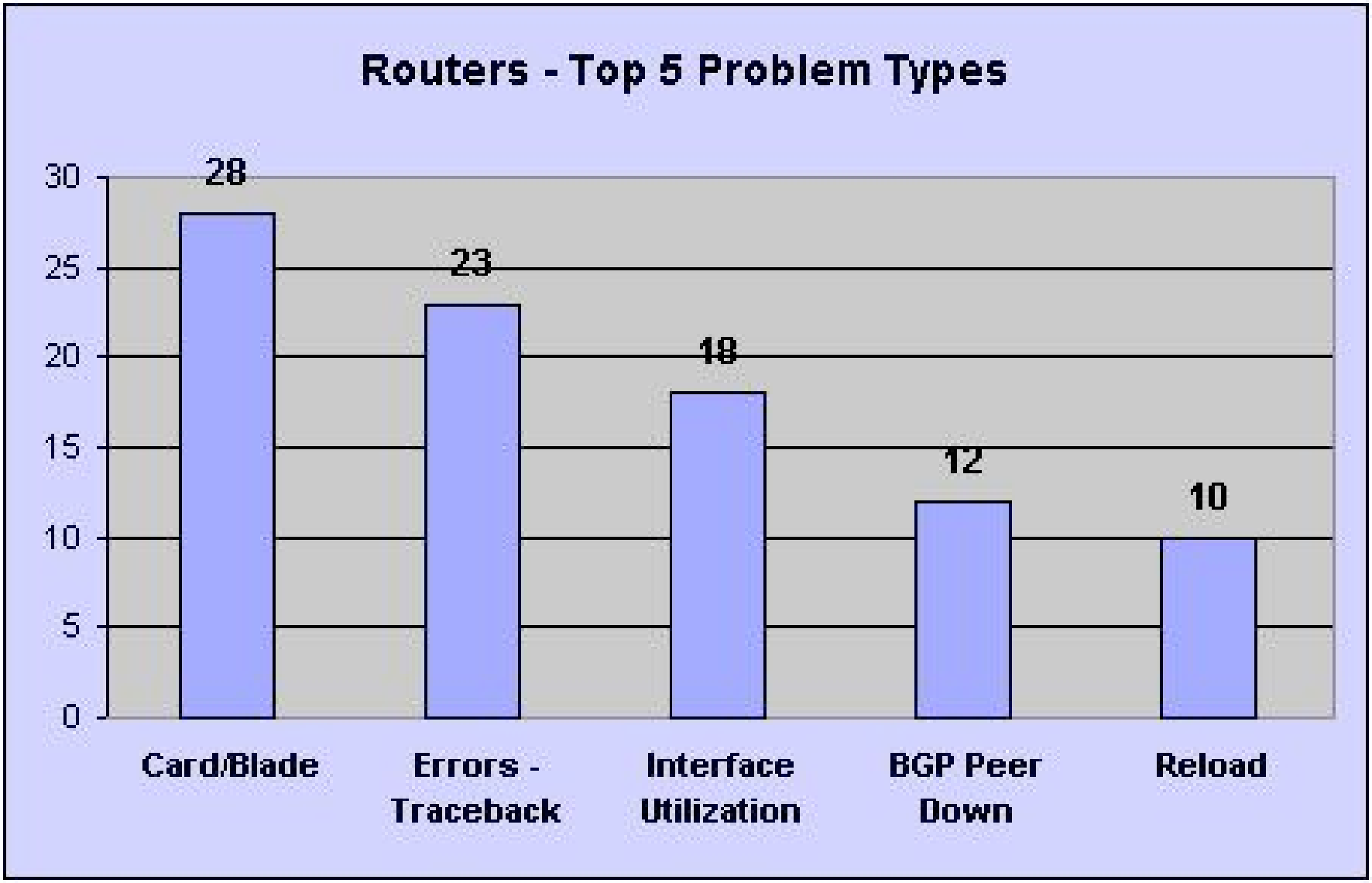


Count (+ maintenance)

NOC/Network	
Problem Area	Count
Router	186
Switch	92
Circuit	88
Internet Routing	54
Host	37
Other	25
Subtotal	481

NOC/Internet Access	
Problem Area	Count
Hardware	123
Software	44
Network	19
Performance	12
Subtotal	198
GRAND TOTAL	679

Top 5 Router Issues



Most Common Prob/Res Types

Most Common Problem and Resolutions Types - - - does not include maintenance tickets

NOC/Network: 373 tickets

Problem	Tickets
Router	152
Switch	92
Circuit	47
Host	41
Internet I	32
Other	9

Problem Type	Count	%
Card/Blade	28	18%
Errors	16	17%
Hard Down	25	53%
Host Unreachable	15	37%
Packet Loss	13	41%
Tools	3	33%

Resolution	Count	%
Replace Equipment	36	24%
Replace Equipment	32	35%
Replace Equipment	7	15%
Host Problem	13	32%
Modify Routing	9	28%
Host Problem	2	22%

NOC/Internet Access - 189 tickets

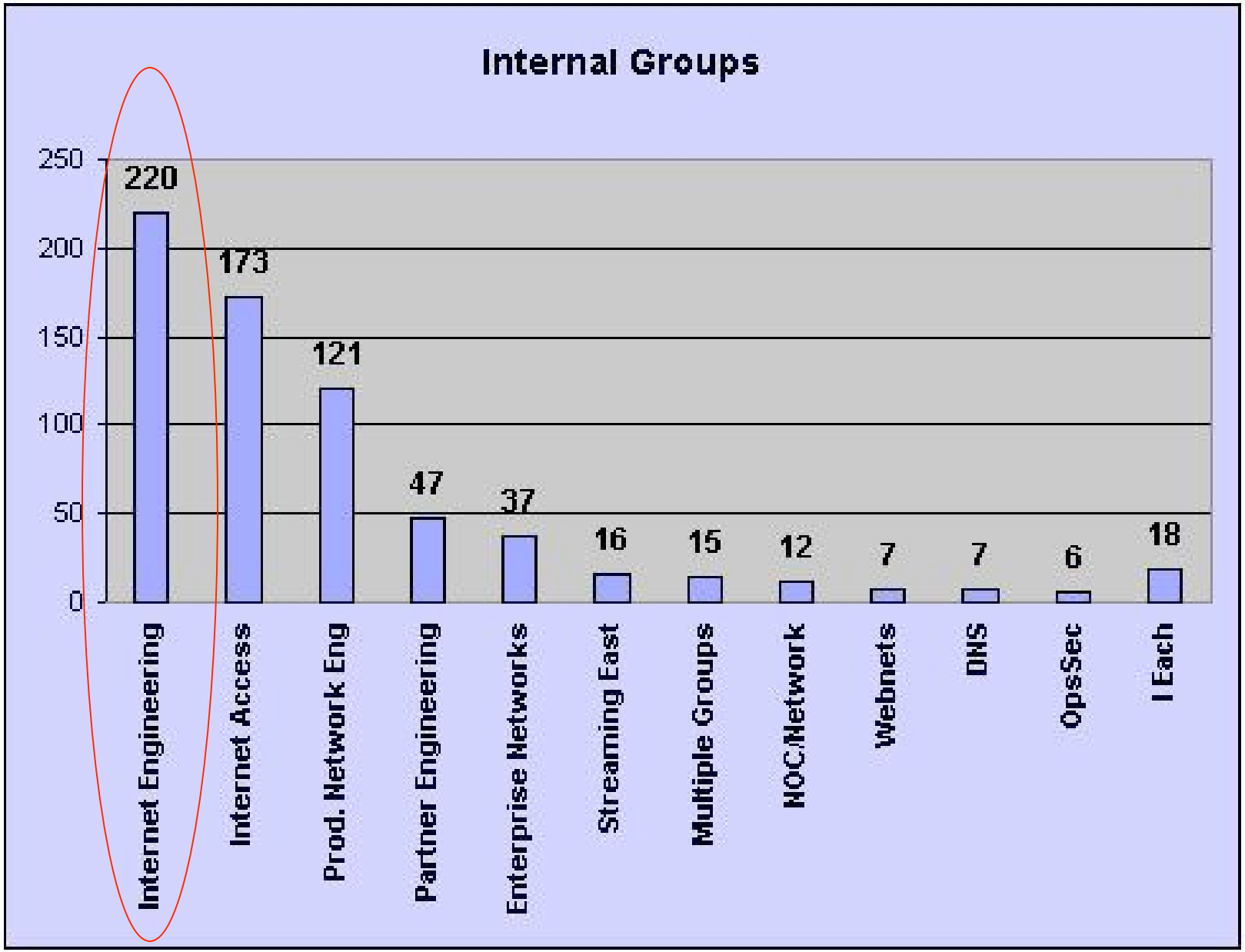
Problem	Tickets
Hardware	118
Software	42
Network	18
Perform	11

Problem Type	Count	%
Host Unreachable	52	44%
File System Capacity	10	24%
Connectivity	7	39%
Streaming	7	64%

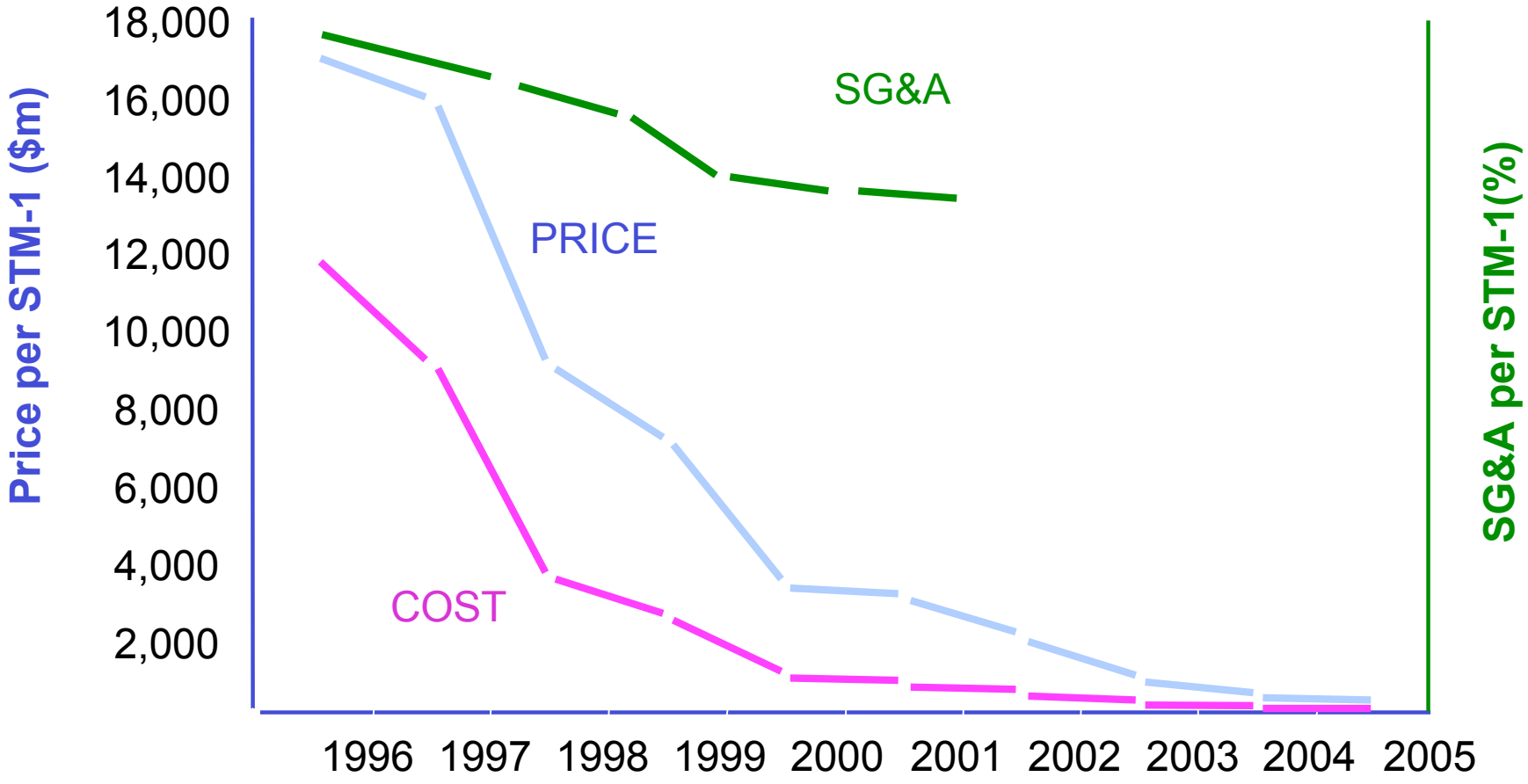
Resolution	Count	%
Reboot	31	26%
Clear file system	11	26%
No action taken	4	22%
No action taken	4	36%

NOTE: All maintenance tickets have been deducted from the above problem and resolution types.

Internal Resources



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

Hardware - General

- Power metering of light-levels on interfaces
  - Very useful operationally
- Protect Flash cards
- Good Stats
  - Need to grab
  - CPU, temperature, fan speed, memory usage
    - High watermarks on queues
      - 5 minute EWMA a random number for microbursts

Control Plane

- Protect the control plane
  - On average, about 2 parity related crashes a day
  - No ECC, no go for ATDN
- Prioritize Hellos
  - Heartbeat hellos
    - Slave takeover from master
  - IGP Hellos over LSAs etc
  - Do not induce further churn at any cost
    - We can live with micro-loops
      - Not so with self-reinforcing oscillatory behavior
    - Exponential backoff on SPF etc.

Hitless Restart

- All Maintenance - Summary
- Normalized problem management tickets

- Type of maintenance
  - Internal : 49%
  - External : 51%

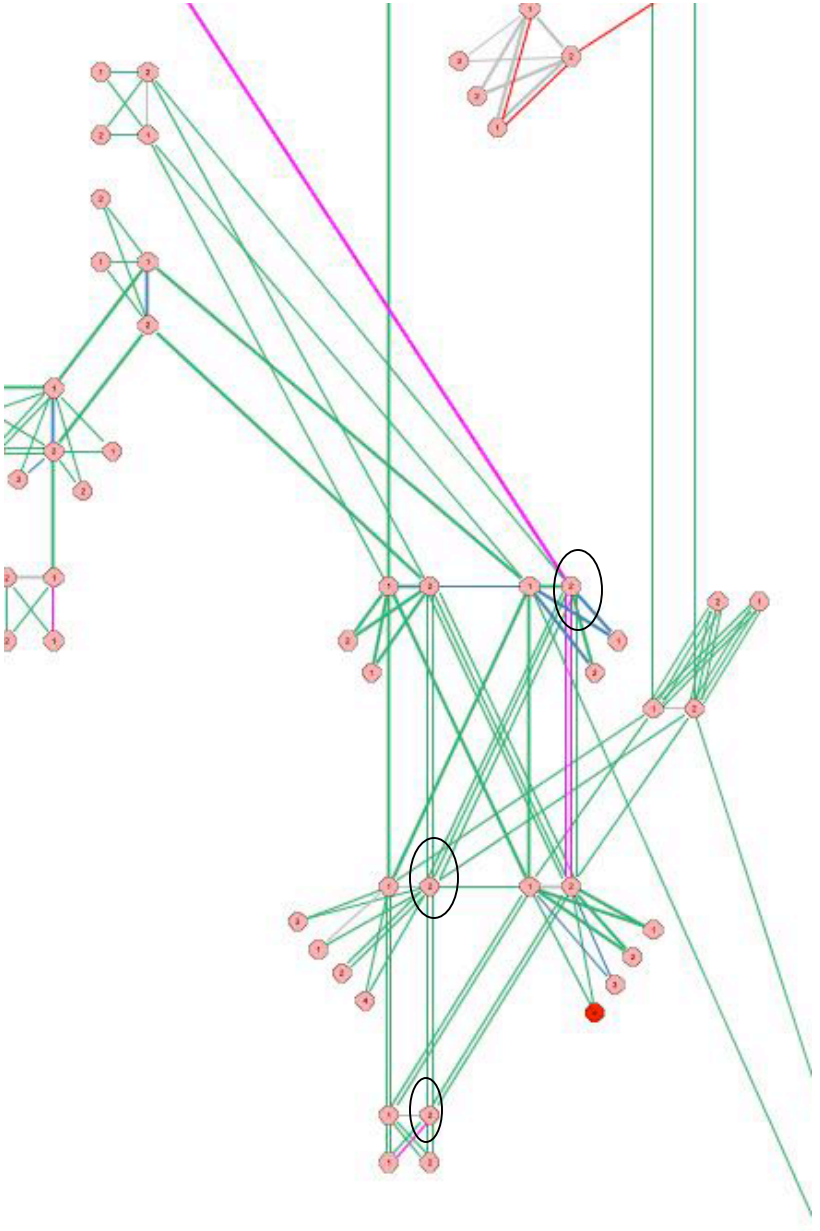
- Impact
  - Completed with impact : 6%
  - Completed without impact : 92%
  - Cancelled : 2%

Software

- Convergence speed
  - Clock turns off when packet forwarding correctly starts
    - Time must include FIB updating
  - Consolidate next-hops
  - Mapping of prefixes to oIF
    - Router or interface crashes
    - Instead of walking FIB to update each individual prefix
      - Update a pointer
- Jitter protocols
  - Introduce fairly large quantities of jitter into routing protocols
    - Update timers, hellos, timed floods etc.
    - Synchronization is bad
      - Don’t particularly care to see timed spikes on routers

ECMP

- Load-share (per flow basis)
- Salt the src/dst/port hash
- Why
  - With a deterministic hashing algorithm
    - Every time traffic is split
    - The hash-space is halved for upstream routers
- Maintenance windows often have near-simultaneous reload of routers
  - Randomly salt



Traffic Matrix

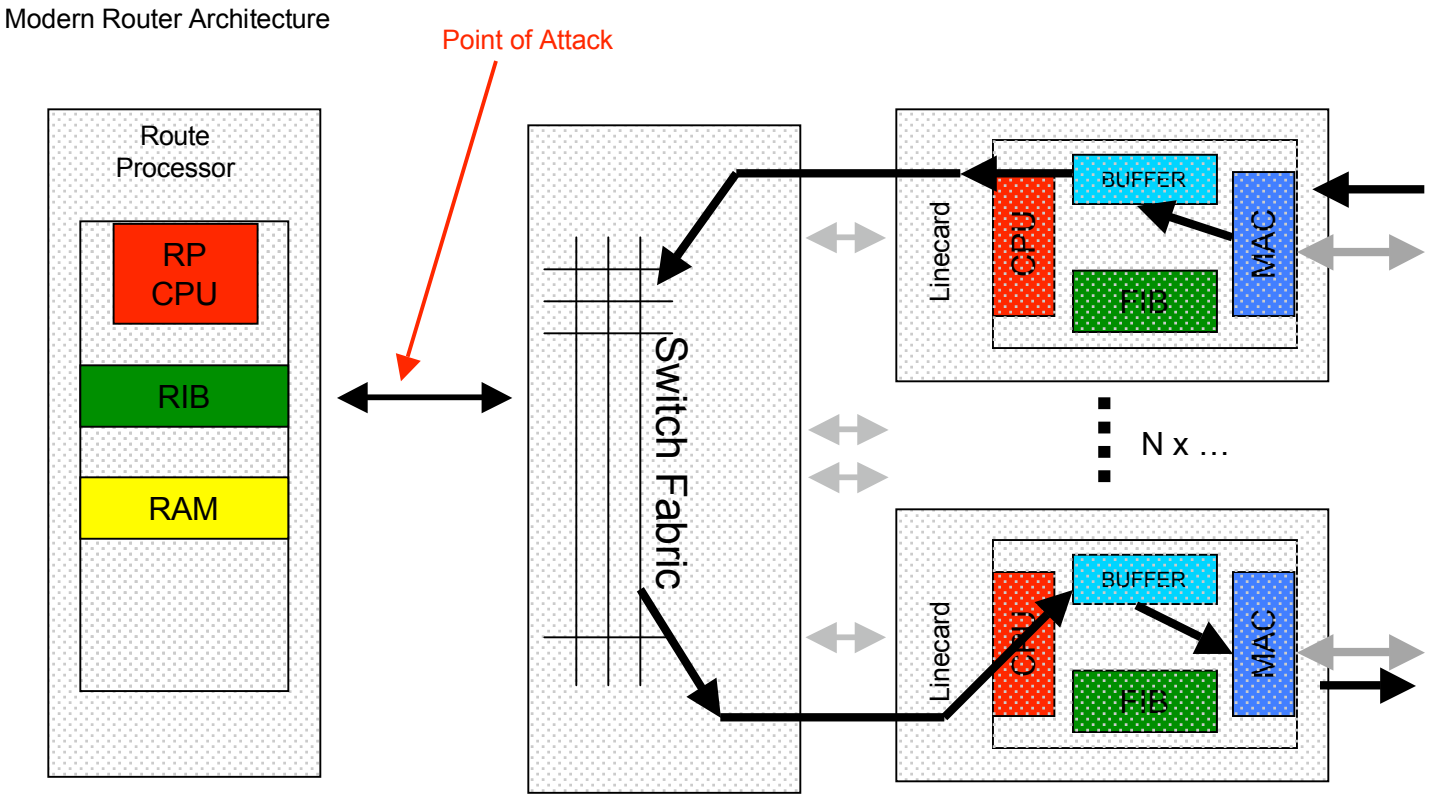
- Proper capacity planning needs good statistics
- Not most vendors strong point
- Flow data interpretation complicated for building POP-POP flows
- Need Router-Router (BGP next-hop) based Flow data

Security and Flow data

- Console IP interfaces should have separate RIB
  - No way to talk to console in-band
- Packet Filtering should work
  - At Line rate
    - At 40 bytes
  - With Flow data
- We use cflowd extensively
- Real output packet filters
  - Inverting and applying all incoming interfaces is not useful

Security

- Routers are optimized for traffic **through** the hardware
  - Not traffic **for** the hardware
- Designing a cost efficient router implies:
  - Cross-sectional bandwidth capacity dominates budget
  - No cost-effective way to engineer a router that can absorb and usefully process data at the rate it can arrive



Hardware – Queuing of Control Plane Traffic

- This one should be easy to get but surprisingly few can do it
- Simple, unambiguous parsing
  - Filter on stuff that is for the router
    - What I deem interesting goes onto the high priority queue
    - Everything else goes onto the low priority queue
- Simple discriminator function/ACL etc.
- Rate-limit on low priority queues
- Apply discriminator on linecard/forwarding engines BEFORE it hits the brain
- Why?

Outside Context Problem

- Attackers are seizing this weak link as a point of attack
  - DoS attacks targeted at infrastructure are increasing
  - Hackers will evolve – Have seen port 179 attacks already (and MSDP can't be far behind)
- Problem
  - Need some way to disambiguate between invalid and valid control traffic (e.g. BGP updates)
  - Rate-limiting on control traffic is not sufficient
    - Enough false data will swamp legitimate data
      - Connection flaps/resets
    - Need to focus on BGP (MSDP)– other traffic is not control, thus will not cause control plane issues

Security

- IGP traffic can be safely blocked
- MD5 on neighbors will not prevent the Router CPU from being inundated with packets that must be processed
- Solution
  - Short term - Dynamic Filtering on the line cards
  - Long term – outboard processing of SHA1/HMAC-MD5
  - This is very long term indeed – not necessarily solving a known problem today (replay or wire sniffing)
  - Vendors have to implement priority queuing for control traffic from line cards to control plane

Dynamic Filtering

- Filtering on the 4-tuple
  - Use the BGP 4-tuple to dynamically build a filter that is executed on the line card or packet forwarding engine
  - Packets destined for the router are matched against the filter
    - If the packet matches the filter
      - Place into the high priority queue
    - Else
      - Place into the low priority queue

Analysis

- On average, will need to try 32000 times to find correct 4-tuple
  - Attacker resources will need to be on average 32000 times greater to adversely affect a router
  - Cost of attacking infrastructure has risen
  - Cost to defender minor
    - Each configured BGP session already has all the state needed above to populate the filter
    - Can use the same solution to protect against MSDP spoofing
- Implementation (sort of)
  - In JunOS (apply-path)

Thoughts

- Stability is most important
  - Only place the high priority queue filter for a neighbor once the session is established
    - Before session is established, place neighbor packets in low priority queue
  - We'll take time for a session to come up over knocking existing sessions down

Thoughts

- Future Goals
  - Use BGP over SSL/TLS (will prevent replay attacks)
    - Can use the filter list along with SSL/TLS to reduce number of valid packets making it to the RP CPU to a comfortable number
- Vendor Feedback
  - Please ensure that your TCP/IP stack chooses randomly when picking a source port (currently most do not)

The BGP TTL Security Hack (BTSH)

- BGP TTL Hack
  - Uses TTL as input into the discriminator
  - http://ietfreport.isoc.org/ids/draft-gill-btsh-00.txt
  - Set TTL to 255
    - Most BGP sessions are between direct neighbors
      - Only allow BGP packets with TTL in 254-255 range
      - Reduces attack diameter dramatically

End

- Questions?
- Acknowledgements
  - Alan Nabors, John Ranalli and the Netops NOC
  - IA, ATDN engineering and coders