

# An ISP Reality Check

# Credits

- Thanks to DC Routing Geeks and other operators for review and correction of early versions of this presentation.
- Reviewers in alphabetical order:
  - Jun-Ichiro Hagino, Joel Halpern, Geoff Huston, Rob Jaeger, Tony Li, Dave Meyer, Mike O’Dell, Dave O’Leary, Andrew Partan, Rob Rockell, Ted Seely, Mike StJohns, Bill Woodcock
- Any presentation bugs are Ran’s fault.

# Limitations

- This presentation describes common and typical ISP properties.
- It is a little US-centric, due to author and reviewer bias.
- This does NOT describe a specific ISP.
- All ISPs vary somewhat from this.
- ISP properties vary over time.
  - In particular, bandwidth keeps getting cheaper

# Outline

- Prime Directive
- Typical Backbones
  - Design, Engineering, Topology, etc.
- Capacity Engineering
- Accounting
- Access Links
- Router Design
- Congestion Avoidance
- Lessons from 9/11/01
- And much more

# Prime Directive

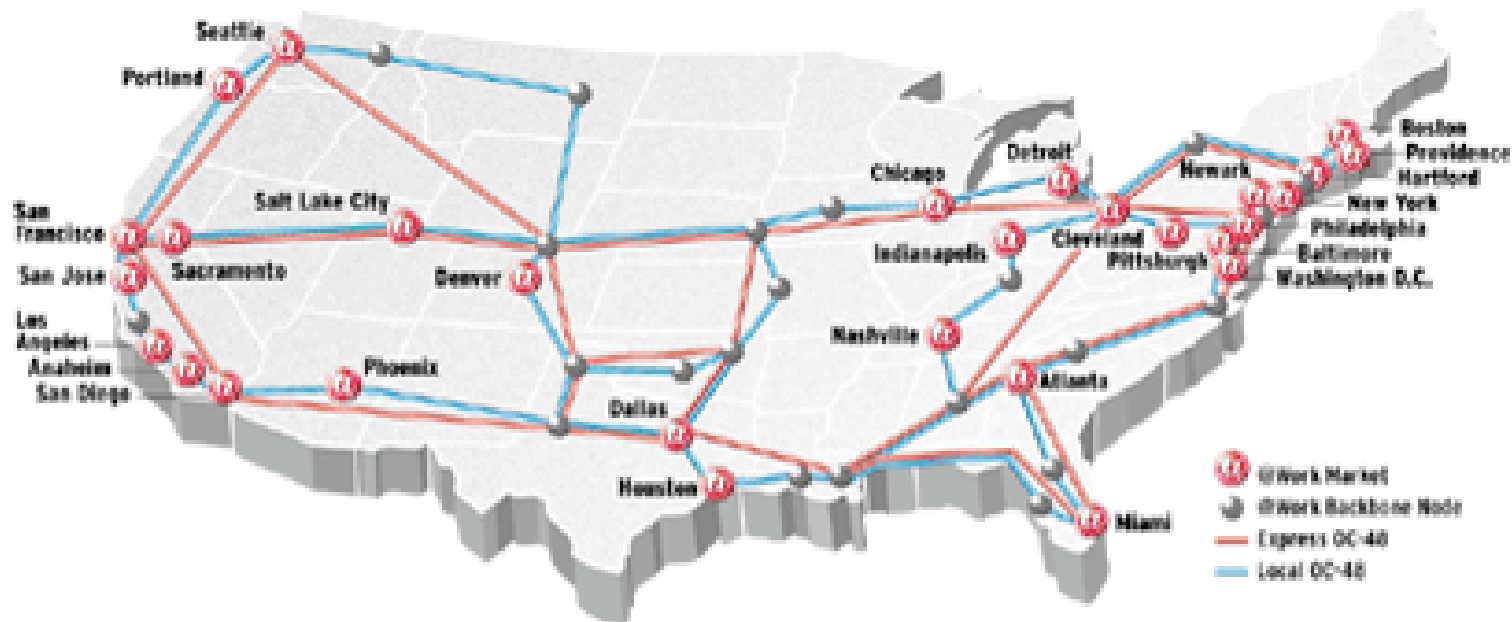
- Drop no packets inside the network, even in the worse case situations
  - Worst case includes: fibre cuts, router problems, etc.
    - Also includes disasters, terrorism, & nuclear war
  - One large multinational ISP reports less than 0.005% packet loss worst case (experimentally measured)
  - Same ISP reports no adverse customer impact despite typically having 2 major fibre cuts/month

# ISP Backbone Characteristics



# Typical Backbone Technology

- PPP over SONET (POS) by far most common
  - Nx OC-12 uncommon outside Asia/Pacific
  - Nx OC-48 common today
  - Nx OC-192 increasingly common, esp in Europe/NA
- ATM over AAL5 (ATM) increasingly rare
  - Nx OC-12 and Nx OC-48 exist in some places
  - OC-192 ATM not available in routers (now or soon)
- Wave Division Multiplexing (WDM)
  - Typically has a SONET/SDH physical interface
  - Permits 40+ Gbps over a single fibre pair

# Former @Home Backbone



- Native access to Excite@Home broadband subscribers
- Nationwide, redundant Tier 1 fiber-optic network
- Scalable 5Gbps, dual OC-48 IP backbone
- State-of-the-art, self-healing technology
- World-class private/public peering

- Cisco Powered Network 
- Server infrastructure by 
- An Excite@Home service



# Capacity Engineering

- Most backbones are over-engineered
  - Because it is lower cost to build them that way
  - Cost of bandwidth keeps dropping with time
  - Avoids packet loss when fibre cuts happen
- Max link utilisation typically 35% of link capacity
  - Permits no packet loss even if multiple fibre cuts
- Upgrade capacity when link has utilisation of 40-60%
- Use high-availability routers
- Deploy routers in a redundant manner

# Exchanging Traffic

- Governed by peering contracts between ISPs
- If pair-wise traffic exceeds 40-100 Mbps, then dedicated link is normally used
  - Either MAN fibre or private fibre at an IX
- Otherwise, traffic usually exchanged over a non-blocking switch fabric at an IX
  - Non-blocking is universally required by IXs

# Packet Accounting

- ISPs usually only offer IP-layer service
- ISPs track basic counters of Interface MIB:
  - Bytes in/out on an interface
  - Packets in/out on an interface
- ISPs do NOT continuously track:
  - Traffic mix on each interface by IP ToS, application, or other upper-layer attributes
- IP-layer deals with packets, not flows or calls

# Issues with QoS Mechanisms

# Quality of Service

- Enabling QoS in a lossless network means the QoS packets are often treated worse
  - Admittedly a counter-intuitive experimental result
  - Particularly true on CPU-based routers
  - Also seen in some ATM switches, by the way
- So far, unable to find any commercial ISP whose eng/ops staff will confirm actual large-scale deployment of IP-layer QoS
  - So far, press releases != reality

# How QoS Increases Costs

- Operations costs:
  - Need to debug “does a packet with <foo> QoS get there” not just “does any packet get there”
  - Need to correctly handle more complex configs
- Harder to troubleshoot whether a routing problem, QoS problem, or both -- hence lengthens MTTR
- Deploying QoS often implies upgrading the deployed hardware
- The more deployed features, the more potential for something to break, hence shortens MTBF and increases operations costs

# Risks of IP-layer QoS

- Source: IEPG meeting, 17 Nov 2002
- Deploying QoS (example: DiffServ) creates a new vulnerability to DoS and DDoS attacks on ISPs -- can reduce service quality
- Large number of edge sites emit priority traffic towards same victim
- Computationally infeasible to authenticate all IP packets with (ToS  $\neq$  0) inside routers

# ISP Services



# Most ISPs Offer Only 1 Service:

- Best-effort IPv4 forwarding

# Access Links

- Commercial Examples
  - DSL,
  - T1, NxT1, T3
  - OC-3 POS/ATM
- Residential Examples
  - Dialup or ISDN
  - DSL
  - Cable Modem
- Primary source of network congestion
- Customer controls capacity of access link
- Customer often controls their router
- ISP can't affect much by itself

Winding Down

# VoIP & Traffic Mix

- VoIP is less than 1% of bytes/packets in a large carmarker's international corporate IP network.
  - Information current as of 11/13/02
- VoIP is a very small percentage of packets/bytes in any commercial ISP
- Growth of other traffic types dwarfs growth in commercial ISPs and corporate networks today

# A Conundrum

- When the network delivers all the packets, it is impossible to provide “preferred” or “lower drop preference” service.
- So, much better to ask that one’s packets get delivered than ask that one’s packets get special treatment.

# Lessons from 9/11/01

- The Internet did NOT have problems
  - Negligible packet loss/congestion in the net
  - Unicast & multicast each worked fine
  - Demonstrated we are ALREADY prepared
- Some content providers had transient problems inside their LANs or servers
  - Generally fixed by moving to no-image content
- Dynamic routing worked VERY well
  - We did have fibre cuts, but routed around them
  - Fast convergence times in modern routing protocols

# Conclusions

- The Internet has demonstrated that it is already prepared for emergencies
  - 9/11/01, many earthquakes/disasters, other history
- Congestion avoidance mechanisms work well
- ISPs design their networks to avoid packet loss
- Congestion is largely an access link phenomenon created by customer choices
- Deploying QoS can reduce IP service quality.

# Impact of Router Design on IP congestion



# Main Points

- Packets do not get lost inside properly designed modern backbone routers
  - Routers have enough buffer for TCP congestion avoidance to kick in and reduce the offered load
- Packets do get lost when the next-hop link lacks bandwidth
  - This happens on access links
  - This does not happen on backbone links; many backbones can even survive fibre cuts without problem
- ISPs make pessimistic deployment assumptions, so they deploy routers in a redundant fashion

# Backbone & Router Design

- Routers are designed to avoid packet loss by facilitating TCP congestion avoidance
- Packet memory on an interface is normally:  
((Interface speed in bits/second) \*  
(trans-Oceanic round-trip-time in seconds))
- Deployed routers have non-blocking switch fabric
- Low-cost WAN fibre --> over-provisioning common

# LAN/MAN & Router Design

- Conceptually similar to backbone routers
- Typical RTT is very very small
  - So less packet memory/interface is needed
  - Ergo, packet memory typically smaller
- Non-blocking switch fabric still common
  - Not all switches provide this; most can
- Availability of cheap 1/10 Gig Ethernet means over-provisioning very common

# Router Forwarding Path

- CPU-based software forwarding for years
- ASIC-based hardware forwarding is now fairly common, but not yet universal
- Properly designed router using ASIC forwarding is generally more robust
- Access routers often still CPU-based:
  - Increased risk of packet loss/congestion
  - More features means lower forwarding rate

# Router Reliability

- Much shorter MTBF than a Class 5 switch
  - Causes ISP to build in router redundancy and lots of fibre path diversity
  - Dynamic routing important to failure recovery
- Improving a lot over time
  - Redundant power common
  - Redundant CPU/switch fabrics common
  - Redundant PHY not unusual
  - More modular software increases robustness

# Congestion Avoidance & Control

- Paper by Van Jacobson, ACM SigComm 1988
  - Defines TCP congestion avoidance algorithms
- TCP-like protocols interpret packet loss as congestion ==> reduce sending rate
  - Sliding window also limits quantity of unACK'd data
- If congestion appears, it generally goes away within about 1 RTT
  - 1 RTT is generally less than trans-CONUS
- IETF also working on ECN

# Service Level Agreements

- Most do not cover the access links
  - ISP can't control that link, so won't make promises
- Most do not cover inter-provider traffic
  - Single ISP can't control whole path, so won't make promises
- Most written by lawyers & accountants
- SLA violation gives user free service for some time in future or maybe a rebate
- SLA Engineering by over-provisioning, not QoS

# ISP Pricing Models Today

- Several pricing models exist today:
  - Flat-rate, but tiered on access link capacity
    - By far this is most common
  - Flat-rate plus usage based on Nth percentile of traffic on the link during previous month
    - Becoming more common; very common for backup
- Other pricing models exist now & in future
- IETF doesn't get involved in pricing models



# Pricing by Access Link Capacity

- ISP costs are generally fixed, not variable
- Fibre or leased capacity costs generally fixed
  - Leased circuit costs vary by link speed
  - Higher speed links lower cost/bit/second
  - Most large ISPs have dark fibre
- Corporate customers prefer flat rate pricing because more predictable/budgetable