GLIF

Akira Kato

Univ. of Tokyo/WIDE Project
kato@wide.ad.jp
Dark Fiber based Networks

☆ National Lambda Rail is operational in U.S.
  - http://www.nlr.net/
  - Dark Fibers from Level3
  - LAX-SFO-SEA-DEN-ORD-PIT-NYC-WAS
  - Southern route is being installed
  - Cisco ONS-15808 DWDM
    - Upto 30 10GE LANPHYS
  - Light Paths can be configured statically

☆ SURFnet operates a nation-wide DF based network
  - http://www.surfnet.nl/
  - Light Paths can be configured statically
Non-DF based Networks

☆ JGN2 in Japan
  • OC-192 or 8*GbE for Long-Haul
  • Some links are 10GE LANPHY based on DF
  • Layer-2 service rather than Light Paths

☆ CA*net4 in Canada
  • Nation-wide Multiple OC-192
  • Only fraction of BW is used for commodity
  • Light Paths can be configured through a GUI
    — UCLP
OC-48/OC-192 class International links

☆ Some are operated as a regular Layer-3 links
  • TransPAC/APAN, SINET, etc
☆ Others are operated as "lambda"
  • some of the bandwidth is reserved for a layer-3 link
  • SURFnet
    — OC-192 between Amsterdam and Chicago
  • IEEAF Atlantic
    — OC-192 between New York and Amsterdam
  • IEEAF Pacific
    — OC-192 between Seattle and Tokyo
  • JGN2
    — OC-192 between Tokyo and Chicago
A Light Path

☆ There is no formal definition yet
  • Generally point-to-point connection w/ QoS enforced

☆ Many levels of definitions can be possible
  • A MPLS tunnel/VPN w/ QoS
  • VLANs
  • ATM PVC w/ CBR or ABR
  • SONET/SDH
  • GbE/10GE encoded on SONET/SDH

☆ Pure Light Path might have
  • No jitter, No packet loss (there may be bit errors)

☆ Layer-2 Light Path
  • Bandwidth guaranteed, some jitter
  • Jumbo frame enabled
  • Transparent to Layer-3 Protocols
Scientific Data Transfer

☆ **Long-distance transmission**
  - Source: Accelerators, Telescopes, ...
  - Destination: Universities, SuperComputer Centers, ...
  - Transmission often across national borders
    - even several times

☆ **Amount of data is huge**
  - Data sources generate lot of data
  - Precision of analysis depends on the amount of data
  - Semi-realtime transmission is required
    - For duration of "observation" or "experiment"
    - May not occupy the bandwidth in 24x7 basis
Scientific Data Transfer

★ **Number of hosts involved is small**
  - 10 hosts in each side, or even single host

★ **Large bandwidth is required**
  - GbE, OC-48, or 10GE/OC-192
  - 100ms or 200ms RTT delay
  - Bandwidth-Delay product is huge
    - Very difficult to control transmission
  - Most of them require reliable transmission
    - VLBI would be an exception
  - Regular TCP is difficult to accommodate it
    - "fairness" among other TCP sessions
    - Aggressive growth of window necessary
    - Even use of other transport protocol than TCP

★ **Jumbo frames required To reduce the overhead**
  - header, interrupt, etc
Global Lambda Integrated Facility

- http://www.glif.is/
- A virtual organization
  - No formal by-law
- Annual invitation-only meetings
  - Aug 2004: Reykjavik, IS
  - Sep 2004: Nottingham, UK
  - Sep 2005: San Diego, US
- Intermediate meeting for Tech WG
  - Feb 2005: Salt Lake City, US
Global Lambda Integrated Facility
Global Lambda Integrated Facility
Sep 2004 : Tokyo -- Geneva OC-192

☆ Multi-national collaboration : JP, US, CA, NL, CH

- 11,000mi/18,000km long
- L2 device: NI40G in T-LEX and CERN
- First long-distance 10GE WANPHY
- The light path existed for two weeks
Data Reservoir tried to fill the pipe
- Prof. Hiraki from the Univ. of Tokyo
- http://data-reservoir.adm.s.u-tokyo.ac.jp/

Each box is with a Chelsio T110 10GE NIC
- TCP was off-loaded to the NIC
- 1500 byte standard frames used

Two different transfer modes tried
- A pair of Opteron boxes
  - Each with a Chelsio T110 10GE NIC
  - Single TCP
  - Memory to memory transfer
- A pair of 9 Xeon boxes
  - Disk to disk transfer
  - Many (36) TCP sessions
  - iSCSI over TCP
Sep 2004 : Tokyo -- Geneva OC-192

☆ **Circuit setup**
  - Configuration has been done in a day

☆ **Packet loss**
  - Divide-and-conquer loopback test
    - Time consuming process
    - SURFnet OC-192 card in Chicago was broken
    - Required another day to replace it
  - Optical attenuator at ONS-15454 OC-192 LR card
    - Special SC connector (with a shutter)
    - Attenuator should be at the other end of patch
  - Clock source issue
    - IEEAF OC-192 doesn’t provide clock
    - It provides just a clear channel
    - T-LEX ONS-15454 got a clock from OC-12
Sep 2004 : Tokyo -- Geneva OC-192

☆ Email was used for communication
  • Good
    — Asynchronous, Recorded
    — Photo, drawing, URL
  • Bad
    — Non-realtime, need a separate trigger
    — SPAM protection can delay the delivery
    — Difficult to tell "subtle" things

☆ May need a sophisticated method
  • International Trouble Ticketing System?
Providing access to boxes helped a lot
- Mutual R/O access provided
  - UW and T-LEX ONS15454’s
- DR got access on NI40G’s in both ends
  - Realtime reading of error counters useful

Making it to public may involve a security issue
- May need to establish a "Trustworth" community
- Access to every gear en-route should be provided
- Abilene’s Core Node Proxy is a wonderful example
- Device independent access possible?
  - ONS-15454, OME6500, HDXc, Foundry, Cisco
  - Force10, Procket, Juniper, Hitachi/Alaxala, Fujitsu
Results

- Circuit was configured w/o packet loss
- 18,000km measured by "known" points
- Single TCP memory to memory
- 7.5Gbps was marked
- Disk to disk transfer
- 9Gbps was marked
DR tried Bandwidth Challenge

• AMS--CHI--TYO--CHI--PIT : 31,248km
• Marked 7.21Gbps Single-TCP memory to memory
• LSR was not fully approved
  – LSR measures by distance of L3 devices
  – AMS--TYO--CHI--PIT : 20,123km

DR’s retry in Christmas

• TYO--CHI--AMS--NYC--CHI--TYO : 30,000km
  – Actual distance was 33,979km
  – But cropped to 30,000km due to a LSR rule
• 7.21Gbps was marked, yielding 216,300 Tbps m/sec
• Updated LSR
  – Single-TCP class
  – Multiple-TCP class
☆ JGN2 : Japan Gigabit Network version2
  - A Nation-wide testbed network
    - Funded by MPHPT
  - Version 1 was ATM based
  - 10GE/GbE based JGN2 launched in Apr 2004
  - http://jgn2.jp/
    - Not all information available
    - Not information available in English
  - (partially unprotected) OC-192 to Chicago

☆ JGN Symposium held twice a year
  - Jan 17-18, 2005 in Osaka
JGN Symposium

☆ Background
• Prof. Smarr from UCSD invited to give a keynote
• He was not make a trip to Japan
• He happened to be in Seattle
• Research Channel at UW get involved
• Trial for remote talk with HDTV
  — Using JGN’s OC-192 circuit to Chicago
• Plan was not discussed in detail by Nov 04

☆ Uncompressed HDTV
• Requires 1.5Gbps bandwidth
• UW’s implementation can stripe it over two GbEs
  — No OC-48 or 10GE is necessary
**Intial Plan**

- Two independent path for redundancy
  - Chicago route for main uncompressed stream
  - IEEAF route for backup HD/MPEG
A fiber cut happened in Jan10 1742JST
• Due to flooding in Nevada
• LA--Chicago JGN2 OC-192 was unprotected
• The carrier was not able to get there until Jan11
Contingency plan was being discussed

- Use IEEAF circuit for two streams
- Even after fiber restored at Jan13 1420JST
  - To avoid confusion
  - UW was busy for three demonstrations
Prof. Smarr gave presentation in almost perfect
• Audio demux broken in the last minutes in Osaka
• Use the combination of
  – Video from the Uncompressed stream
  – Audio from HD/MPEG
• Lip-Sync was lost, unfortunately

For those who involved to make this successful
• UW, Pacific-Interface, PNWGpop, NLW, StarLight
• TransPAC, APAN, IEEAF, T-LEX, WIDE, JGN2
• NTT Communications, NTT Lab, NTT West
• and more
Control Plane of GLIF

☆ Most of the gear in the middle
  • Configured manually
  • Coordination by meeting required
  • Actual configuration is done in a day or two
    – "wiring" a fiber may be required
  • The configuration lasts for a few days or more

☆ CANARIE developed UCLP
  • User-Controlled Light Path
  • Configuration through a GUI w/ authentication
  • Generate TL1 commands to be uploaded to gear
  • "time-share" in order of an hour possible
    – one app for daytime, another for nighttime
Control Plane of GLIF

- GMPLS is not easy to introduce
- A Light Path may spawn
  - Multiple administration domains
    - Authentication and Authorisation required
    - May skip accounting in this case
  - Multiple vendors gear in multiple layers
    - DWDM, TDM, OXC
    - Encoding interface: GbE into STS-24
    - Layer-2 devices, VLANs
Summary of GLIF

☆ It is working
  • Very static, or PVC-like in other words
  • A light path lasts for hours, days, or even weeks
  • Reconfiguration may need a day or two
  • Control Plane Protocol
    — Email, Phone, or buy a glass of beer
  • JP, US, CA, UK, NL, CH, CZ, KR, TW
  • End-to-end light path was proved useful

☆ Single Lambda was GbE/STS-24
  • Current trends of single lambda is 10GE/STS-192
  • More bandwidth required
    — 3--4 10G between Japan to Seattle in SC2005
    — DR may fill the pipe w/ PCI-X 2.0