Internet2 Land Speed Records in the GLIF environment

Akira Kato, Ph.D

The Univ. of Tokyo/WIDE Project
kato@wide.ad.jp

Background

☆ **We knew thousands of TCPs can fill a 10GE pipe**
- This was what we have seen in Internet

☆ **How about a single TCP could?**
- This was a challenge
  - Very minor packet drops affect the performance
  - Especially true in large RTT environment
  - TCP algorithm is one of the challenges
- Machine architecture could affect it
  - CPU, BUS, NIC, ...
- Rackful machines are not convenient to hand-carry
  - Single machine in each side is the ideal

A Trigger

☆ **In July 2004, APAN Meeting in Cairns, AU**

☆ **Rene Hatem told me:**

- "Are you interested to have a 10GE lightpath between Tokyo and Europe?"

☆ **I responded immediately:**

- "Certainly!!"
- The tentative configuration:
  - TYO -- SEA -- NYC -- AMS -- GVA

☆ **It was a great idea to have a >20,000km pipe**

- But it was not interesting just to have a pipe
- Somebody needs to fill the pipe with bits
- I talked Prof. Hiraki who happened to be there
  - "Are you interesting to get involved?"
- He responded promptly:
  - "Needless to say!"

Data Reservoir Project

http://data-reservoir.adm.s.u-tokyo.ac.jp/

☆ **A research project**

- Chaired by Prof. Hiraki, The University of Tokyo
- Funded by JP Government

☆ **The basic idea:**

- Develop a set of boxes
- Put the scientific data to one of them
- They transfer the data efficiently over long distance
- Then the scientists can get the data locally
- "It was stupid to force scientists to learn TCP"
- "They should concentrate on their jobs"

☆ **DR participated SuperComputing Conferences**

- 2002 (Baltimore)
  - "Most Efficient Use of Available Bandwidth Award"

Prof. Hiraki in Univ. of Tokyo

☆ **A professor at Univ. of Tokyo**
- Worked on a dataflow machines at ETL before

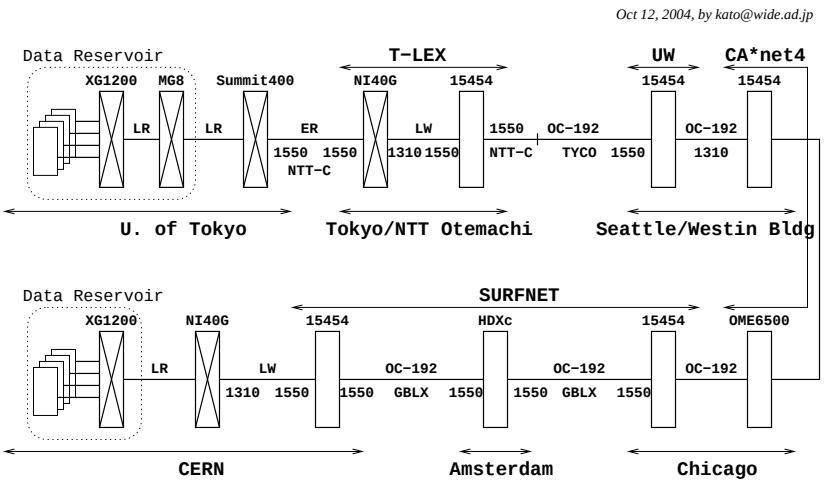
☆ **Computer Architecture expert**
- He was interested in filling up 10Gbps pipe in terms of the computer architecture

The planning

☆ **GLIF Nottingham Meeting in Sep 2004**
- All the parties concerned was there
  - PNWGIGApop, CANARIE, SURFnet/NetherLight, CERN
- A luncheon meeting was held
  - Which circuits were to be used
  - What configuration were to be done
  - When, how, ...
- There was no formal "procedure" to setup a lightpath"
  - Contact info of each network/exchange collected

Very first trial

☆ Oct 2004 Configuration : 11043mile/17772km



** Detailed configuration inside of CA*net4 cloud is not shown*

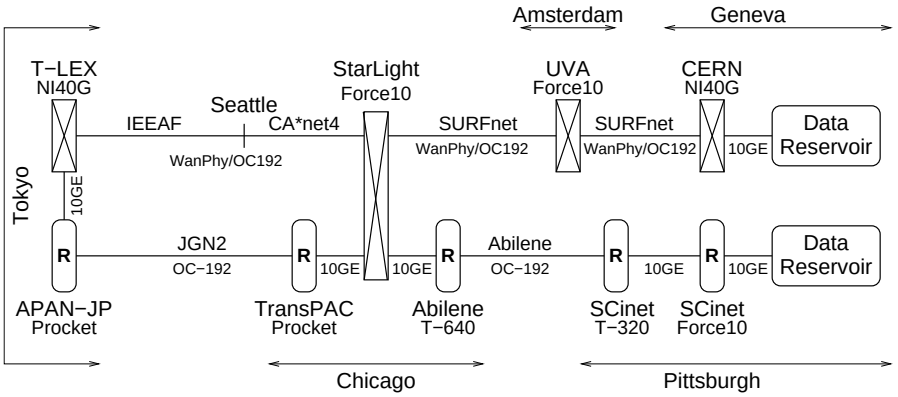
** HND−PDX−SEA−YYJ−YVR−YYC−YQR−YWG−MSP−ORD−AMS−GVA: 11,569mi/18,618km*

Setting up the lightpath

☆ It required almost one week

- Hardware failure
  - Overnight delivery of spare blade
- Communications done by Email
  - Good for recording
  - Bad for realtimeness
- Time difference
  - No common "working" hours among Asia/America/Europe
  - Single transaction could take half a day
- No "lightpath" debugging tool
  - Loopback request via Email was the tool
  - Put a loopback to narrow the section by half
- Subtle tricks
  - Attenuator vs ONS15454 OC192-LR
- BI8000 didn't work well with WANPHY XENPAK
  - Replace it with a NI40G

The second trial in SC04



☆ **Lightpath setup has been done in a few days**

Lessons learnt

☆ **Posting detailed configuration helps a lot**
- NDL helps a lot for this purpose currently
- Updating the description is always important

☆ **Remote loopback manipulation helps**
- Can be done via TL1 proxy
- Password protection and authorization required

☆ **Working in 2am JST works well**
- Everybody else is in the office

☆ **It was not subject for LSR**
- Layer-3 points in Oct 2004
  - Tokyo and Geneva only : 9,816km
- Layer-3 points in Nov 2004 (SC04)
  - Geneva, Tokyo, Chicago, and Pittsburgh : 20,645 km

Key rules of I2 LSR

http://lsr.internet2.edu/

☆ **Four classes**
- IPv4 or IPv6
- Single stream or multiple streams

☆ **Evaluated as "performance * distance"**

☆ **Distance measured by L3 points**
- No L1/L2 point is evaluated
- Maximum distance is 30,000km

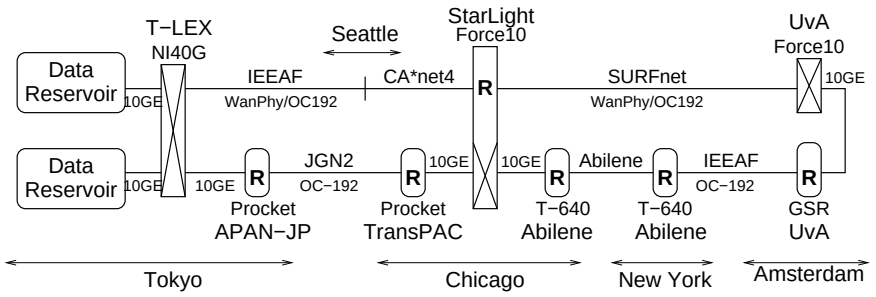
☆ **Need to include "operational" network**

☆ **Need to improve at least 10% of the previous one**

☆ **End system need to be purchasable in the market**
- No special hand-crafted NIC allowed

Another trial (Revenge)

☆ **2004 Christmas Holidays**
- Nobody uses lightpath
- Abilene link utilization was at minimum



The progress

☆ **Nov 9, 2004 (SC04)**
- 7.21Gbps, 148.9Pbm/s, IPv4 single (20,645km)

☆ **Dec 24, 2004**
- 7.21Gbps, 216.3Pbm/s, IPv4 single/multiple

☆ **Oct 28, 2005**
- 5.94Gbps, 91.8Pbm/s, IPv6 single/multiple (15,461km)

☆ **Oct 29, 2005**
- 5.58Gbps, 167.4Pbm/s, IPv6 single/multiple

☆ **Nov 10, 2005 (SC05)**
- 7.99Gbps, 239.8Pbm/s, IPv4 single/multiple

☆ **Nov 13, 2005 (SC05)**
- 6.22Gbps, 185.4Pbm/s, IPv6 single/multiple

☆ **Nov 14, 2005**
- 6.96Gbps, 208.8Pbm/s, IPv6 single/multiple

The progress (cont)

☆ **Feb 20, 2006**
- 8.80Gbps, 264.1Pbm/s, IPv4 single/multiple

☆ **Dec 30, 2006**
- 7.67Gbps, 230.1Pbm/s, IPv6 single/multiple

☆ **Dec 31, 2006**
- 9.08Gbps, 272.4Pbm/s, IPv6 single/multiple

☆ **In summary**
- Variants of Linux were used
- 4 IPv4 LSRs and 6 IPv6 LSRs
- Most of them are >30,000km path
- All of them are with single TCP session
- 9.08Gbps is the last LSR in OC-192c age
  - 9.988Gbps required to beat it

The Trophies

The machines

☆ **In early stage, Dual Opteron was used**
- Better memory access latency
- Chelsio T110 on PCI-X
  - IPv4 TCP/IP off-loading
  - IPv6 TCP/IP off-loading not available
- Chelsio N110 on PCI-X
- No jumbo support is required

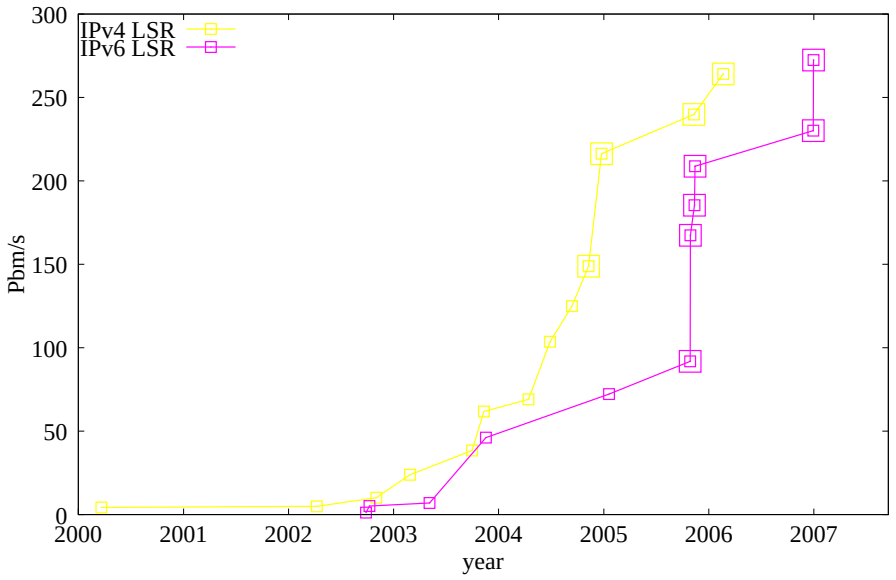
☆ **In later stage, Woodcrest Xeon was used**
- Nice CPU performance
- PCI-X 2.0 based NIC cards
  - Neterion Xframe II
  - Chelsio T310-X
  - Chelsio S310E
- GSO (Generic Segmentation Offload) was used
  - Checksum calculation was offloaded as well

Lessons learnt (end system)

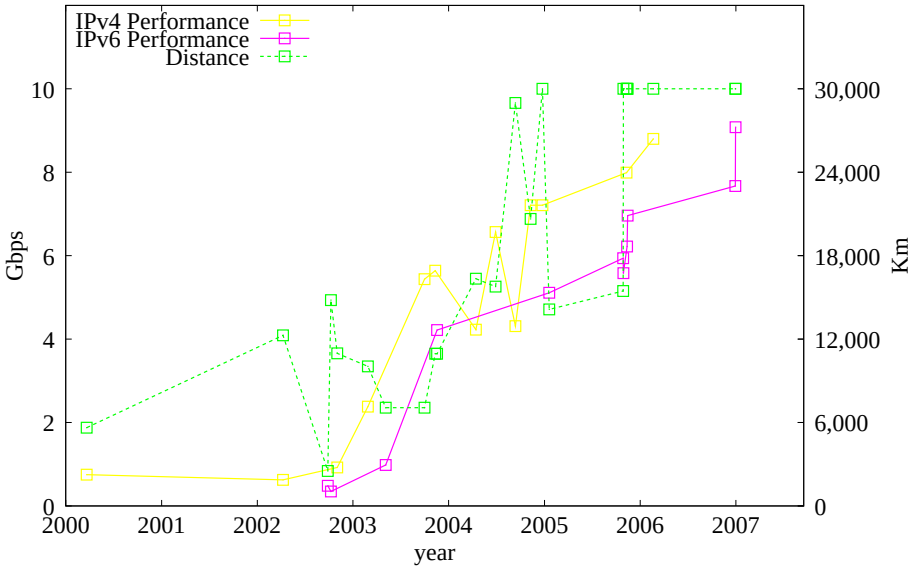
☆ **CPUs fully utilized to process packets at 10Gbps**

- A delay box can't emulate the real network
- Minor things could yield packet drops
  - "cron"
  - Jitter generated by routers/switches
  - It is affected by mode of operation (i.e. L2 or L3)
- FPGA based packet monitors works well
- Sender-side pacing is required
  - Everybody can understand in advance
- Receiver-side pacing also works well
  - Minimize the jitter at receiver side
- Pacing was performed in a FPGA based box
- Tuning for pacing rate was required
  - Manual configuration
  - No automatic method established

Growth of LSR



Growth of LSR (performance and distance)



Growth of LSR

☆ **Two major factors contributed LSR very much**
- Especially after 2004

☆ **10GE NICs**
- Available since later 2003
- Before then, GbE was the forefront

☆ **GLIF's contribution**
- OC-192c's have been common since 2004
- GLIF's international collaboration contributed a lot
- Minimized the L2/L3 devices on the route

Considerations

☆ **LSRs were just for memory-to-memory copy**
- They were useless for production purpose now
- Disk-to-disk copy is at least required
  - Can a single TCP stream fill the pipe?

☆ **Layer-2/3 devices might generate jitter**
- Its extent depends on
  - Manufacturer and model
  - Cross traffic and other functions on the device
- Pacing on both of sender/receiver effective
- Pure L1 lambda reduce jitter

☆ **LSR trials lasted for up to a few hours**
- Can they run in sustained manner?
  - What happens on a residual error?
- How is reproducability?
  - LSRs depended on manually tuned parameters

Conclusion

☆ **Data Reservoir team won 10 LSRs**
- 4 for IPv4, 6 for IPv6
- up to 9.08Gbps single TCP stream

☆ **LSRs were not only done by DR**
- Many GLIF participants and GOLEs
- Concept of GLIF
- Victory of entire GLIF community!

☆ **When lightpaths are used for production purposes**
- We need to provide professional support
  - Many users are not specialist on networking
- Fault isolation and debugging methodology is required
  - When lightpath becomes unusable
  - When the quality of lightpath degraded

→ **Still we need to work hard together...**

Acknowledgement

☆ **On behalf of Data Reservoir Team,**

I’d like to thank all GLIF commuity

for support of all LSRs.