A single-server, single stream disk service on LFN

Data Rerervoir project
The University of Tokyo

What is Data Reservoir?

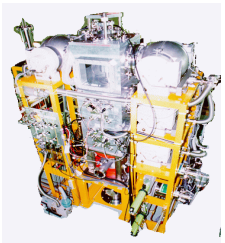
- Sharing Scientific Data between distant research institutes
  - Physics, astronomy, earth science, simulation data

- Very High-speed single file transfer on Long Fat pipe Networ
- High utilization of available bandwidth

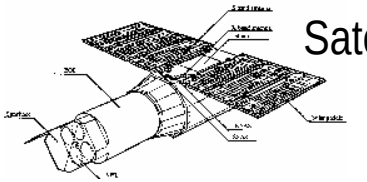
- OS and File system transparency
  - Storage level data sharing
  - Fast single file transfer

Application fields of Data Reservoir

Nuclear experiments

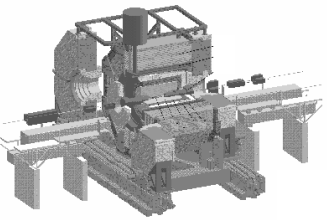


X-ray astronomy
Satellite ASUKA



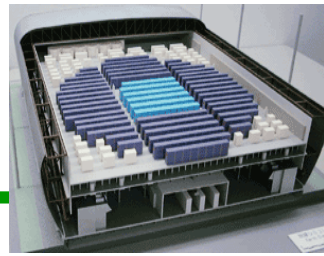
SUBARU
Telescope

Belle Experiments



Data
Reservoir

Very High-speed
Network



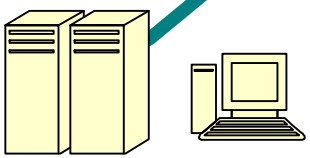
Earth
Simulator

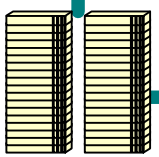
Data
Reservoir

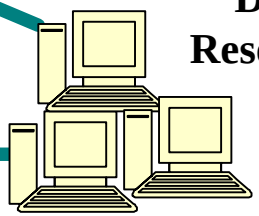
Distributed
Shared
files

Data
Reservoir

Local
Accesses







Grape6

Data analysis at University of Tokyo

Installed Data Reservoir system(2003 ～)

University of Tokyo

Server room

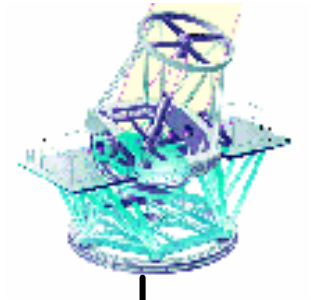
National Astronomica lObservatory Japan

JAXA

Univ. of Tokyo

Univ. of Tokyo

Basic Architecture



Data Reservoir

High latency
Very high bandwidth
Network

Disk-block level
Parallel and
Multi-stream transfer

Local file accesses

Cache Disks

Data Reservoir

Distribute Shared Data
(DSM like architecture)

Local file accesses

Cache Disks

Challenge of 2006, Disk to Disk data transfer

- the key to wide application areas
  - Web, Grid, HDTV, Global storage
- Memory to memory TCP data transfer (2004-2005)
  - 8.9Gbps IPv4 and 7Gbps IPv6 TCP single stream
  - Internet2 Land Speed Record (2004, 2005、2006)

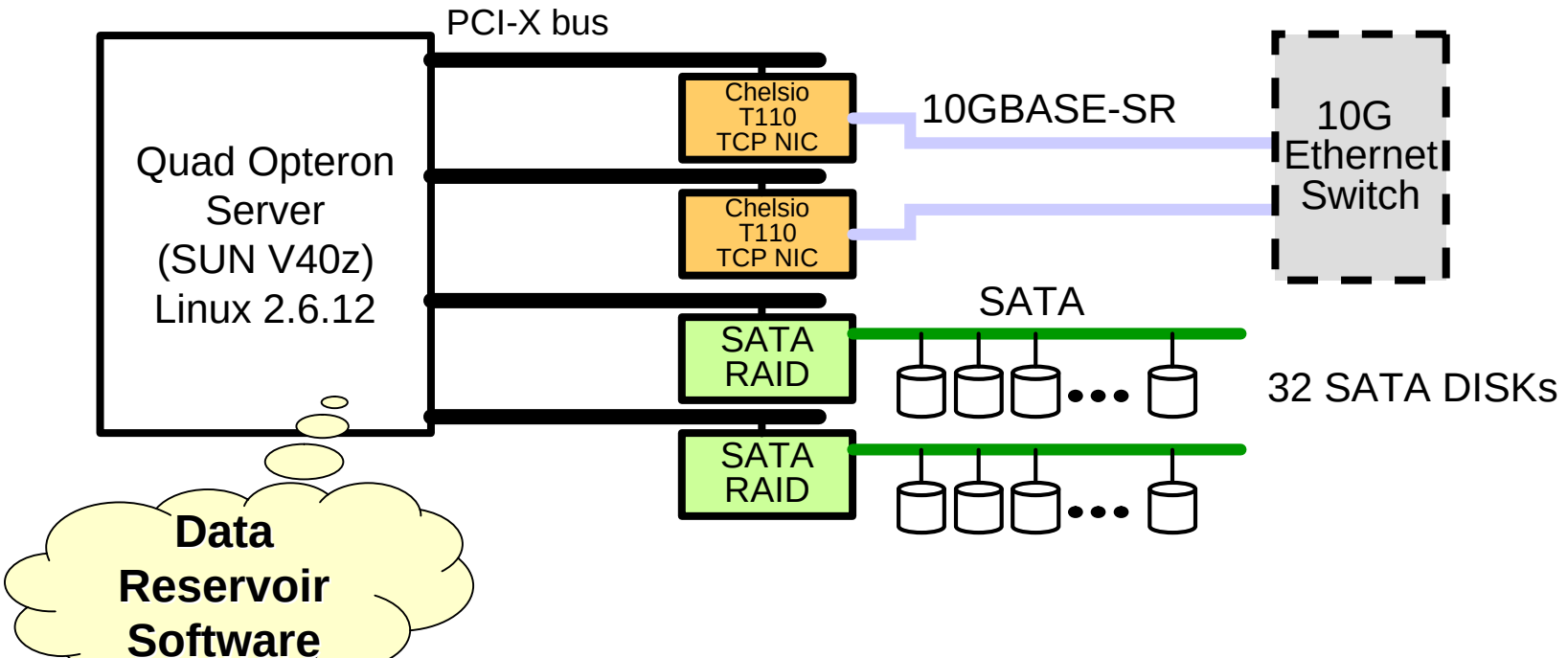
- Disk to disk TCP/iSCSI data transfer (2006)
  - Single stream 7.2Gbps (PCI-X bus bottleneck)
  - Dual streams 8.65 Gbps (93%of peak BW)
- Single server system with 32 SATA disks

Disk data
7.2Gbps single stream
8.65Gbps dual streams

- Same performance onlong distance and local

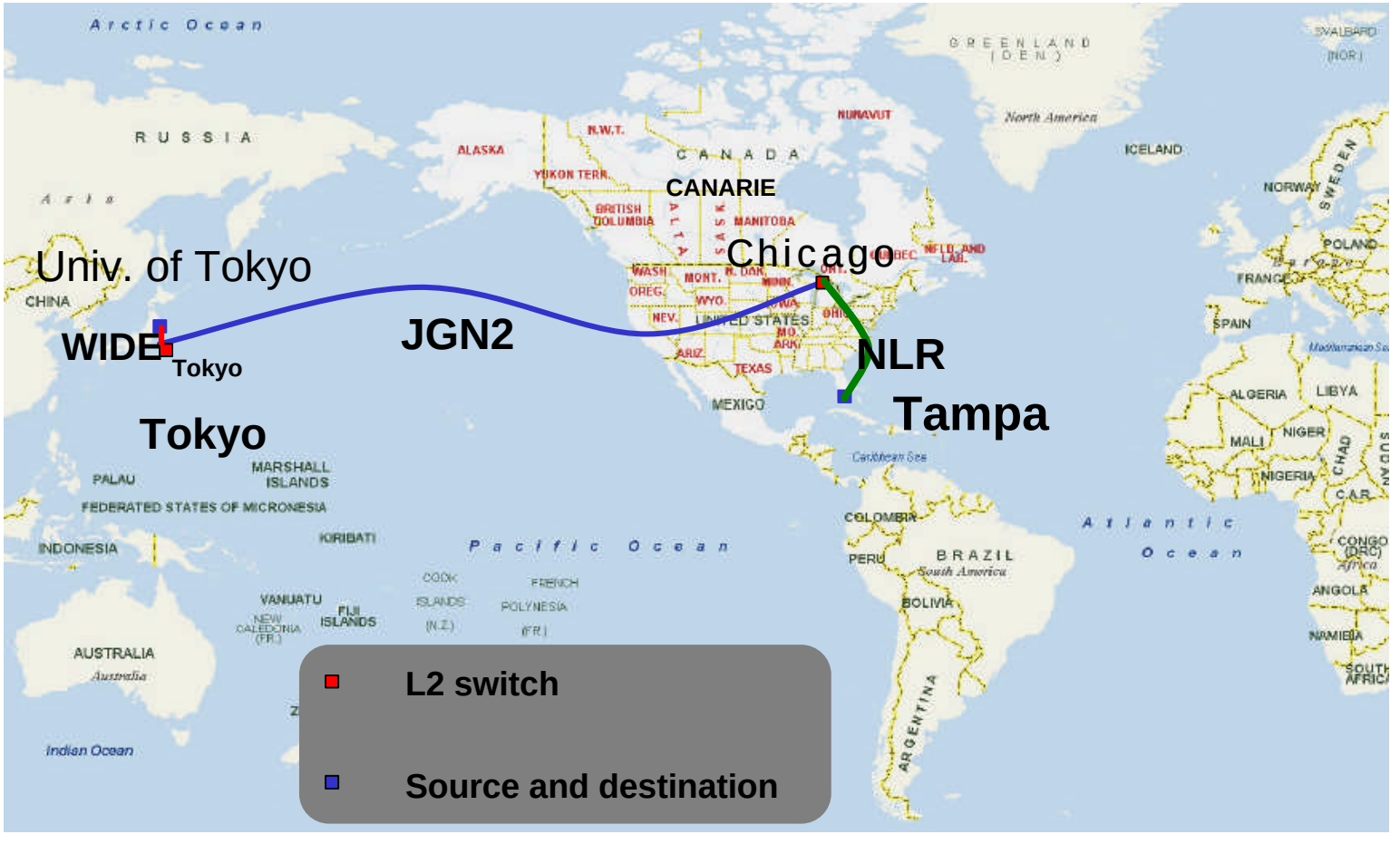
Single box, dual stream Data reservoir

- **A single box 10 Gbps Data Reservoir server**
  - Quad Opteron server with multiple PCI-X buses
  - Two Chelsio T110 TCP off-loading NIC
  - 32 SATA Disk arrays
  - Data Reservoir software (iSCSI deamon, disk driver, data transfer maneger)



Network paths

- Tampa --- Tokyo 11,775Km
- RTT 206 ms



Switch arrangement for BWC measurement

E600
UTokyo

Measurement devices

MG8
UTokyo

RX-4
NEZU
WIDE

NI40G
Tokyo
T-LEX

GS4000
Tokyo
JGN2

CAT6509
Washington
NLR

CAT6509
Pitts
NLR

CAT6509
Cleveland
NLR

CAT6509
Chicago
NLR

E1200
Chicago
StarLight

GS4000
Chicago
JGN2

E300
LA
JGN2

CAT6509
Raleigh
NLR

CAT6509
Atlanta
NLR

CAT6509
Jacksonville
NLR

CAT6509
Tampa
NLR

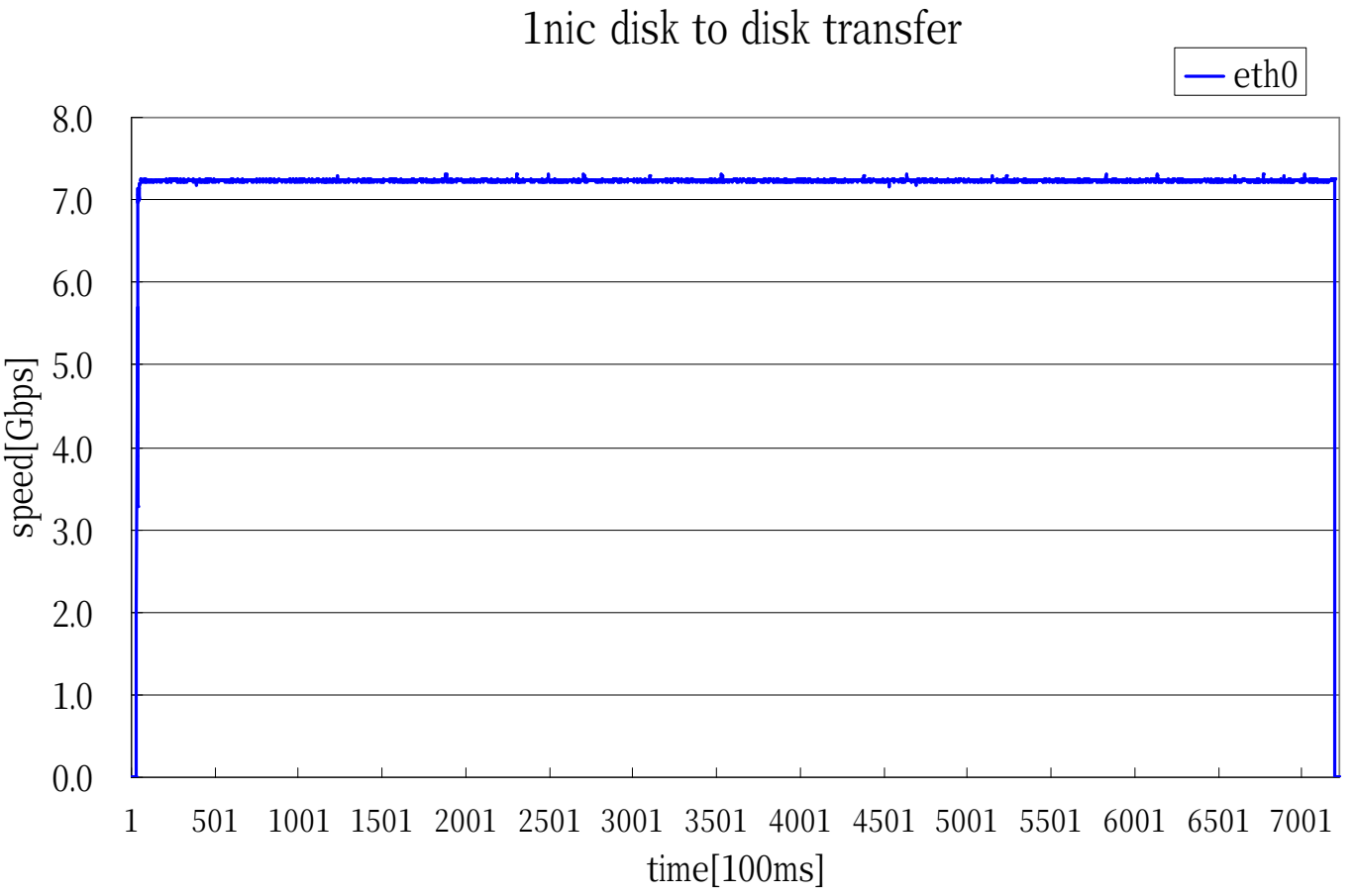
Measurement devices

RX-4
Tampa
UTokyo

RX-4
Tampa
UTokyo

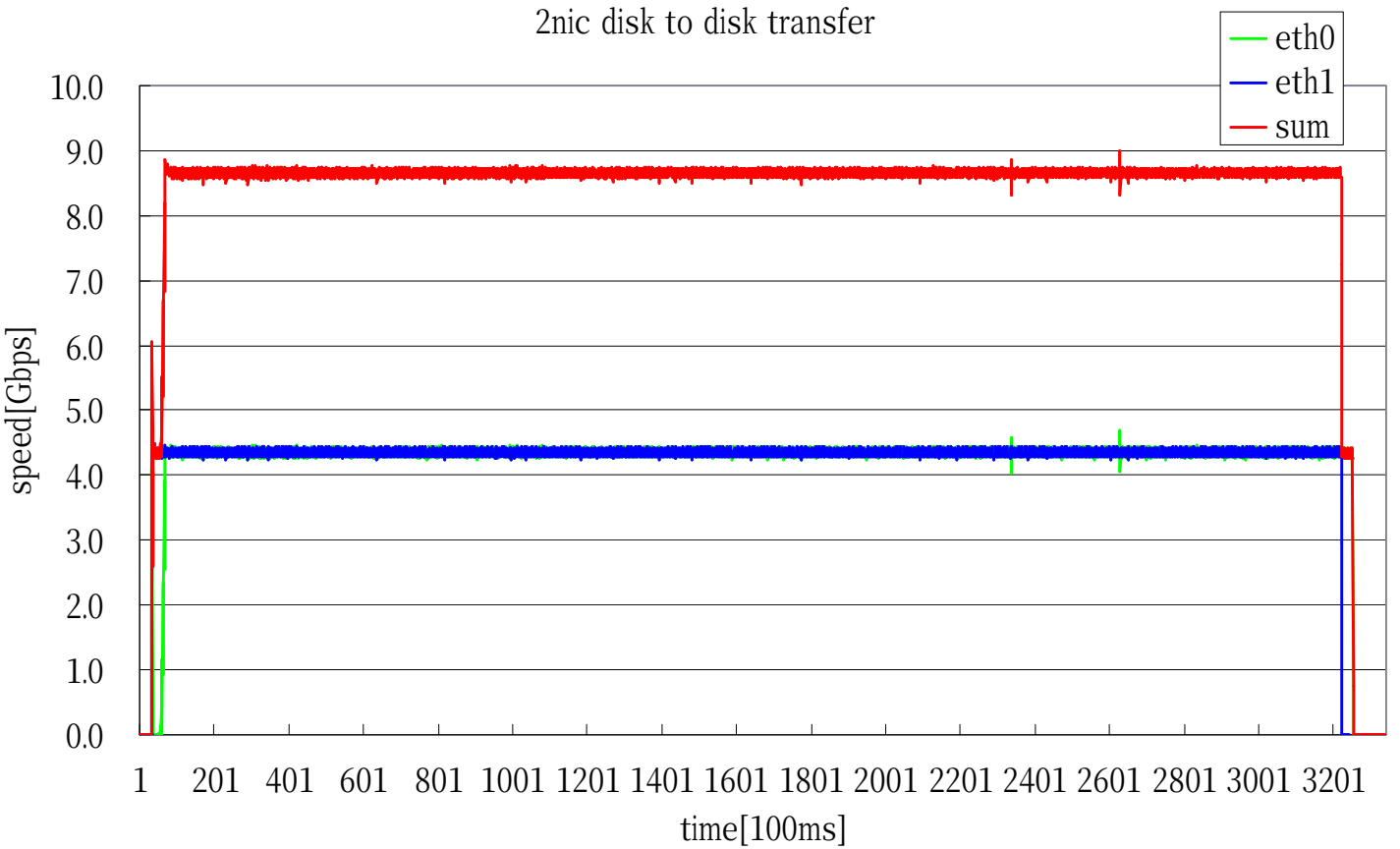
Test results at Tampa

- Single Stream 7.2 Gbps Tampa -> Tokyo
  - 78 % bandwidth utilization (PCI-Bus bottleneck)



Test results at Tampa

- Dual Stream 8.6 Gbps Tampa -> Tokyo (using two NIC of a server)
  - Heavy pacing by NIC and TGNLE-1 (external pacing box)
  - 93.4 % bandwidth utilization



The challenge

Efficient 10 Gbps disk to disk data transfer on iSCSI

- Single TCP stream
- Single server with 32 RAID SATA disks
- SC06 ⇔ the University of Tokyo

- Contribution

(1) First single-stream 10G disk server

(2) First single box 10G disk server

(3) Same performance between local and 11,775Km

(4) Ready to practical use at research institutes

Data Reservoir: BWC teem members

- The University of Tokyo
  - Kei Hiraki, Mary Inaba, Junji Tamatsukuri, Yutaka Sugawara, Takeshi Yoshino
- Wide project
  - Akira Kato, Katsuyuki Hasebe
- Fujitsu Computer Technologies
  - Masakazu Sakamoto, Takuya Kurihara, Ryutaro Kurusu, Yukichi Ikuta