

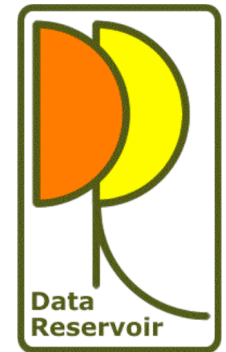
Data Reservoir : **Data sharing facility for** **Scientific Research** **- Hardware approach -**

Kei Hiraki

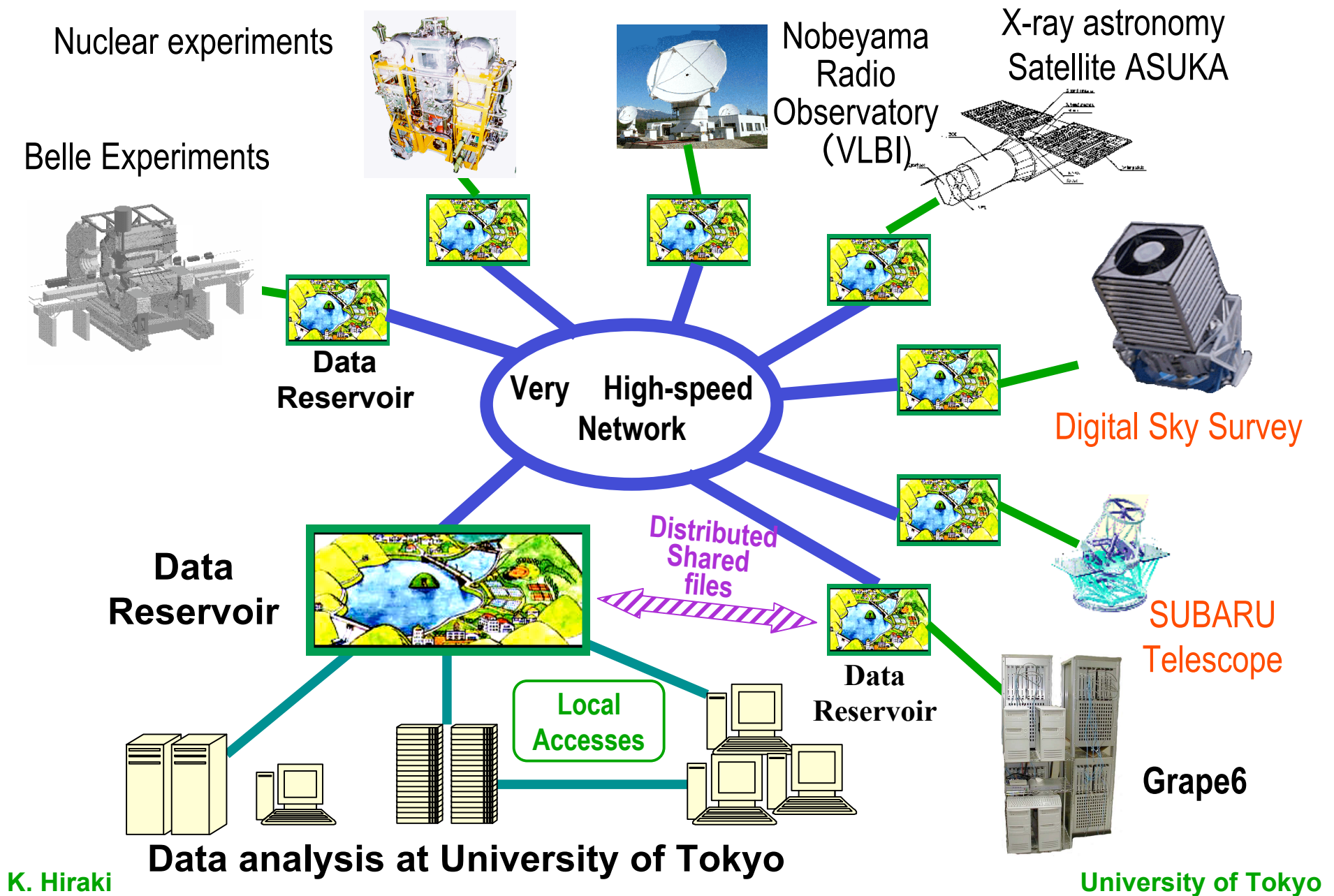
University of Tokyo

Fujitsu Laboratories

Fujitsu Computer Technologies



Data intensive scientific computation through global networks



Research Projects with Data Reservoir

Name	Project	Domestic Connection	Oversea Connection	Current amount of traffic
Hideyuki Sakai	High Energy Polarimeter SMART	RIKEN RCNP	CERN Brook Haven	CERN LEP 70 DAT/month CERN LHC 100 MB/sec Brook Haven 100 Mbps RCNP Accelerator 50 GB/day
Yoshiaki Sobue	Radio telescop (VLBI)	Nobeyama Radio Observatory	Max Plank Observatory	VLBI data ? @ 100 GB
Sadanori Okamura	Slone Digital Sky Survey	National Astronomical Observatory	Fermi Lab.	Survey Data ? F ? @ 10 TB Data Exchange between Fermi Lab
Kazuo Makishima	Satellite observation of early universe	ISAS Hiroshima Univ. Saitama Univ.	NASA European Space Agency	Current Satellite 1GB/day
Toshio Yamagata	Simulation of Global Change	Frontier Research System for Global Change	N/A	, Simulation ? @ 10 TB Currently, data archive system with 50 PBytes
Tomio Kobayashi	JC ATLAS Experiment	KEK Kyoto Univ. Univ. of Tsukuba	CERN	CERN LHC 100 MB/sec
Takashi Onaka	Infra-red observation Satellite	IRIS Nagoya Univ	ESA receiving site ? Sweden ? j	Downlink ? @ 200 MB Data exchange within a minutes
Jun'ichiro Makino	Astronomical Simulation by GRAPE-6	National Astronomical Observatory	Advanced Study, Princeton Univ. Musium of Natural History	Maxmum Throughpu ? F 100MB/s 1 Simulation ? @ 10 TB
Hiroaki Aihara	KEK ? @ , , ? , † , ? , f , " , ? , ' , ™	KEK Naboya Univ.	Princeton Univ.	Raw Data ? @ 100 GB/day Data exchange ? F ? @ 10 TB/day
Sadanori Okamura	SUBARU telescope	National Astronomical Observatory	Hawaii Observatory	100 GB/day ? B Peak bandwidth 0.5 GB /sec ? i , Gbps)

Dream Computing System for real Scientists

- Fast CPU, huge memory and disks, good graphics
 - Cluster technology, DSM technology, Graphics processors
 - Grid technology
- Very fast remote file accesses
 - Global file system, data parallel file systems, Replication facilities
- Transparency to local computation
 - No complex middleware, no or small modification to existing software

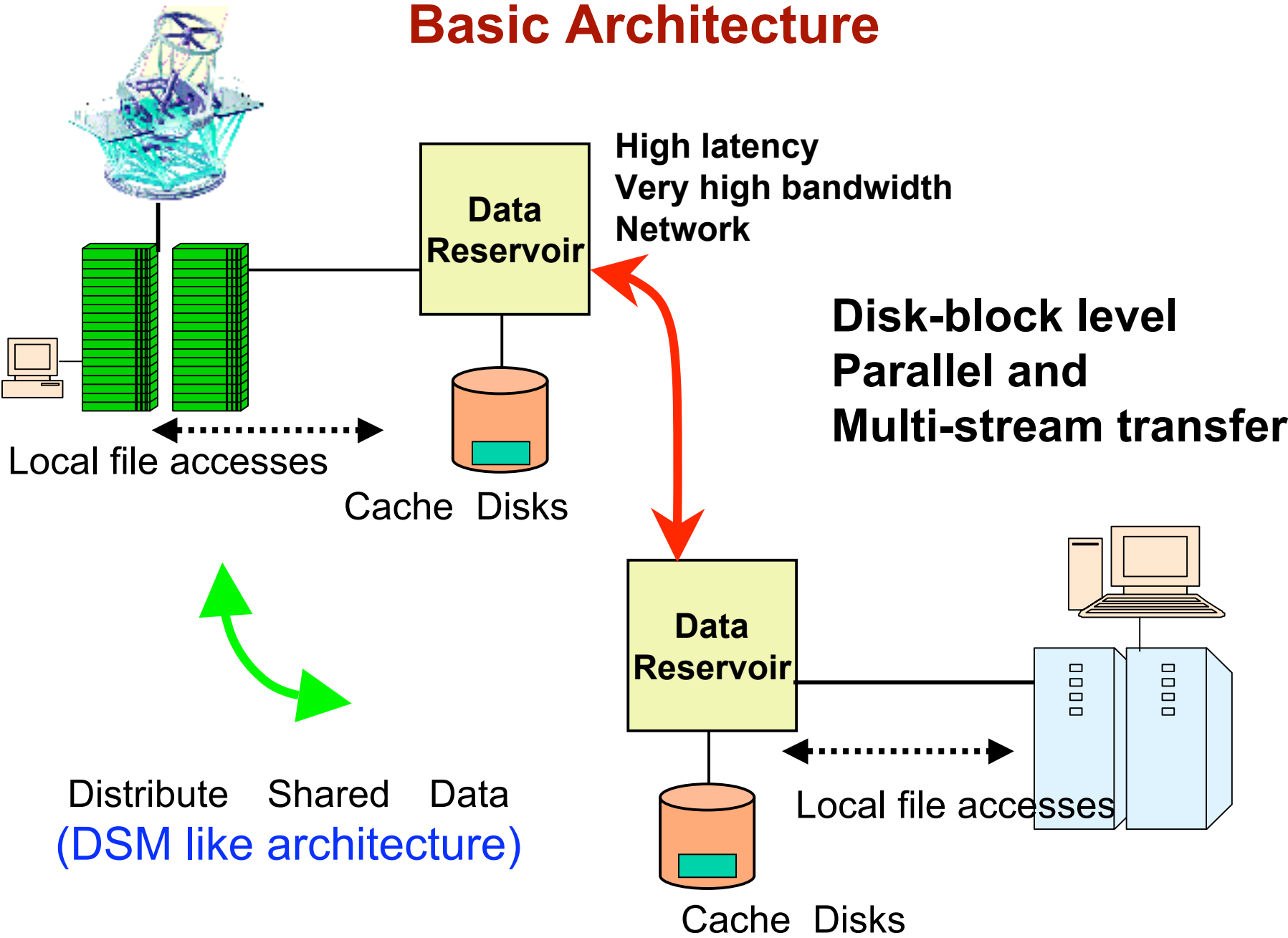


- **Real Scientists are not computer scientists**
- **Computer scientists are not work forces for real scientists**

Objectives of 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 Network
 - > 10 Gbps, > 20,000 Km (12,500 miles), > 400ms RTT
- High utilization of available bandwidth
 - Transferred file data rate > 90% of available bandwidth
 - Including header overheads, initial negotiation overheads
- OS and File system transparency
 - Storage level data sharing (high speed iSCSI protocol on stock TCP)
 - Fast single file transfer

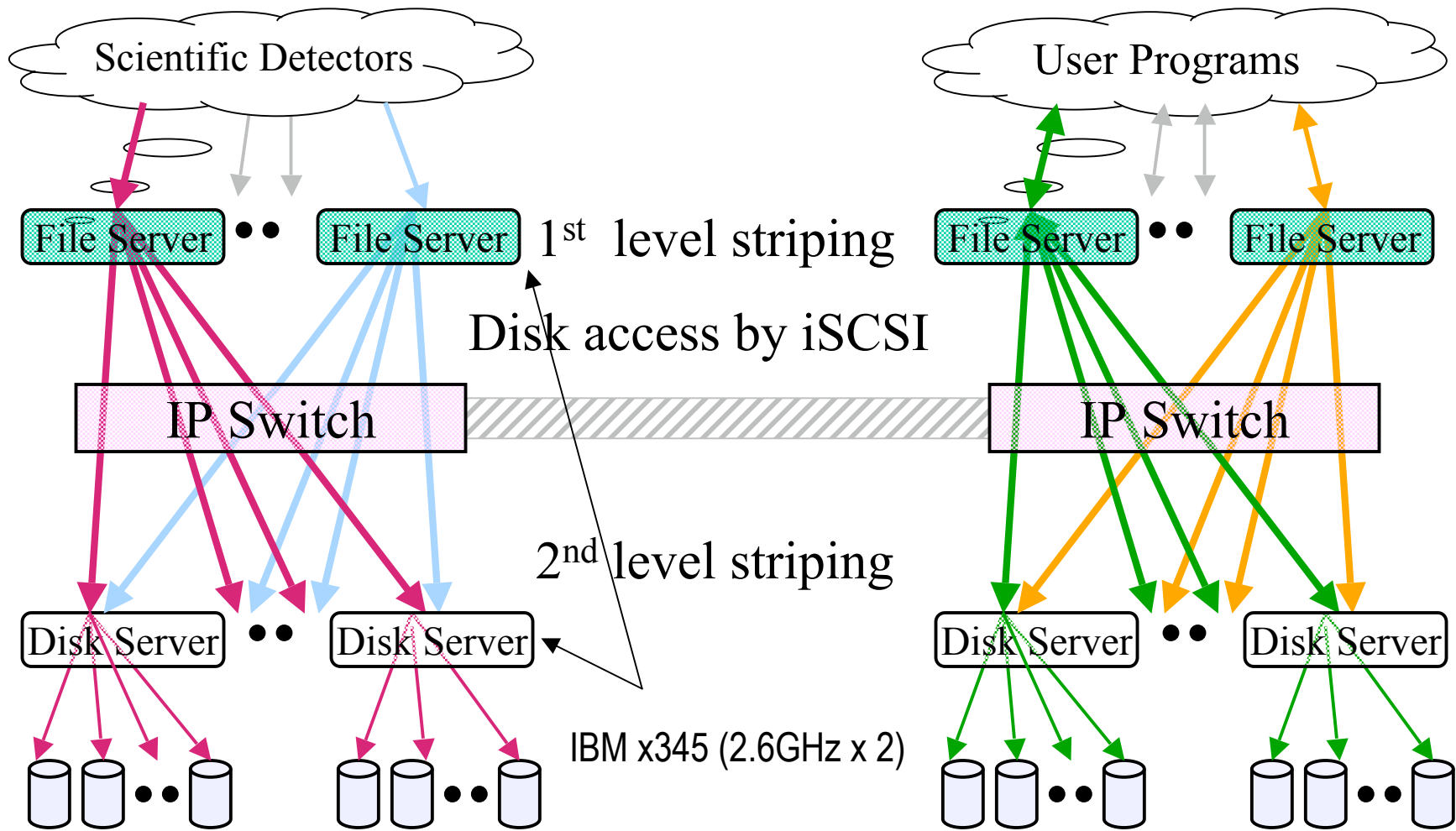
Basic Architecture



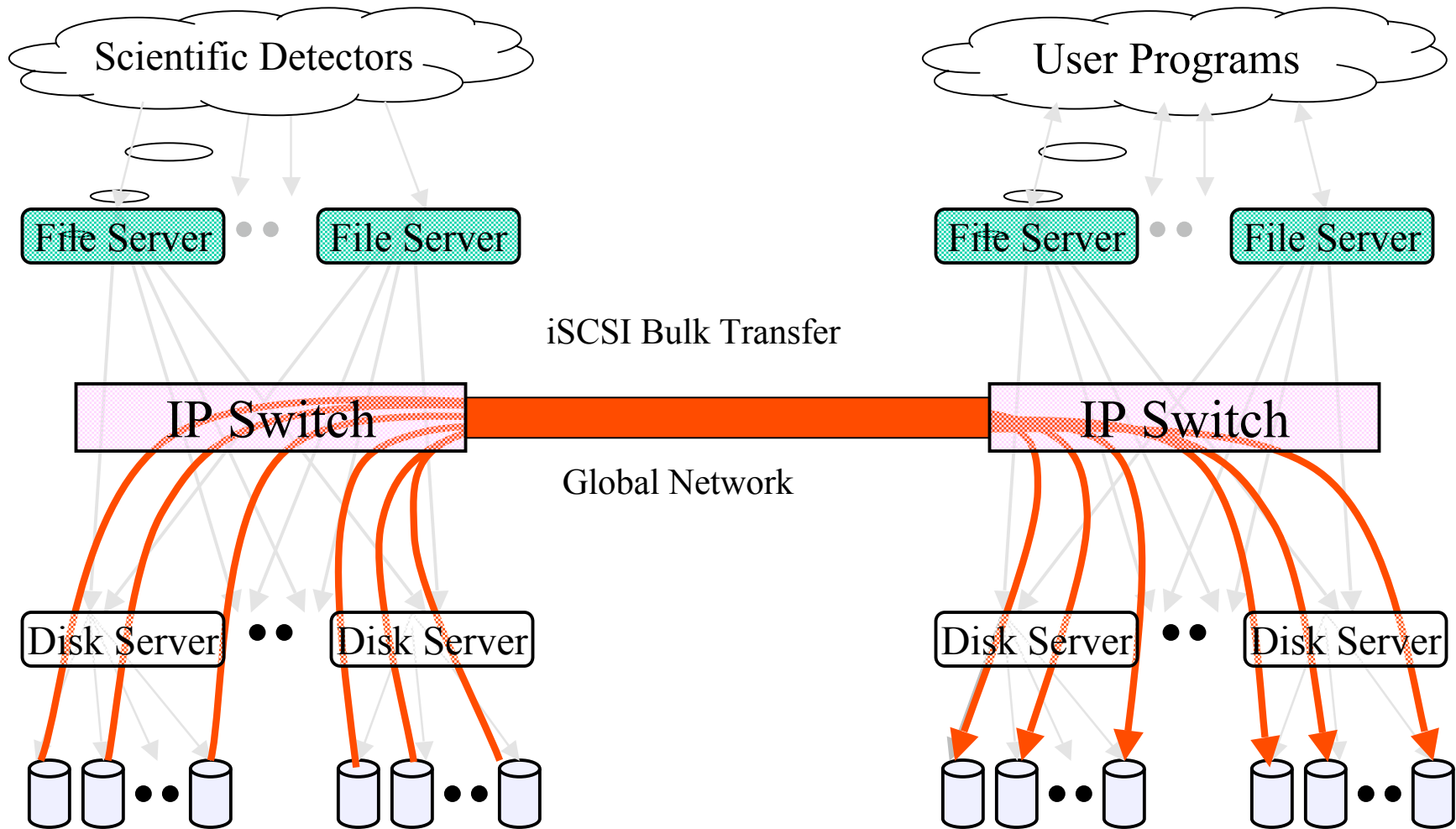
Data Reservoir Features

- Data sharing in low-level protocol
 - Use of iSCSI protocol
 - Efficient data transfer (optimization of disk head movements)
 - File system transparency
 - Single file image
 - Multi-level striping for performance scalability
 - Local file accesses through LAN
 - Global disk transfer through WAN
- } Unified by iSCSI protocol

File accesses on Data Reservoir

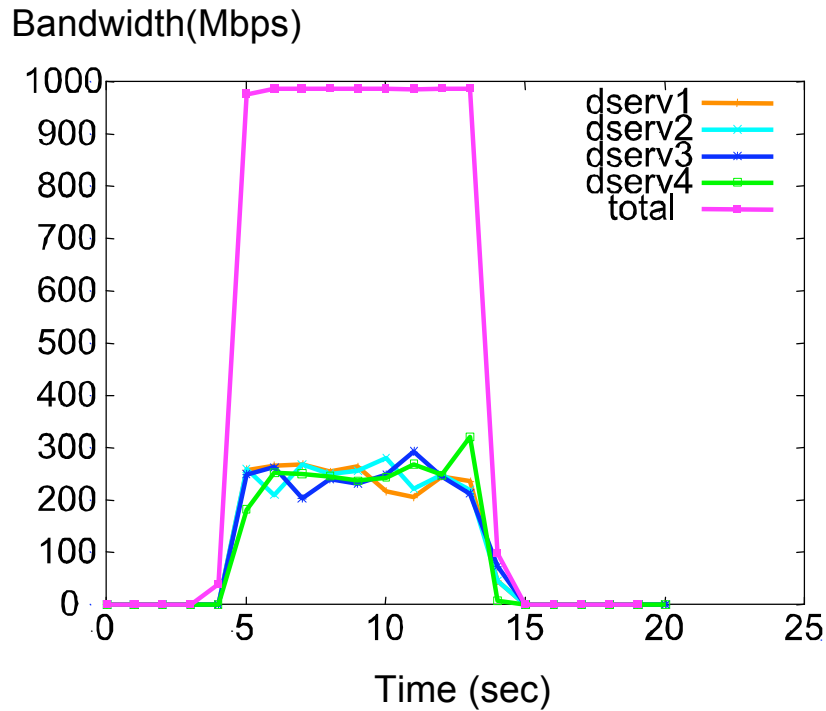


Global Data Transfer

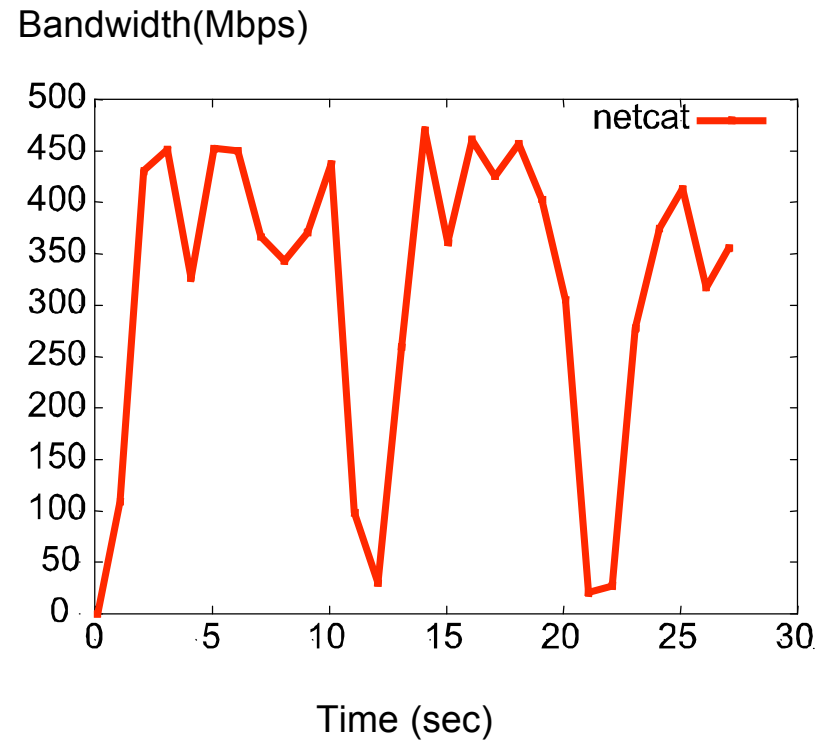


BW behavior

Data Reservoir



Transfer through A file system

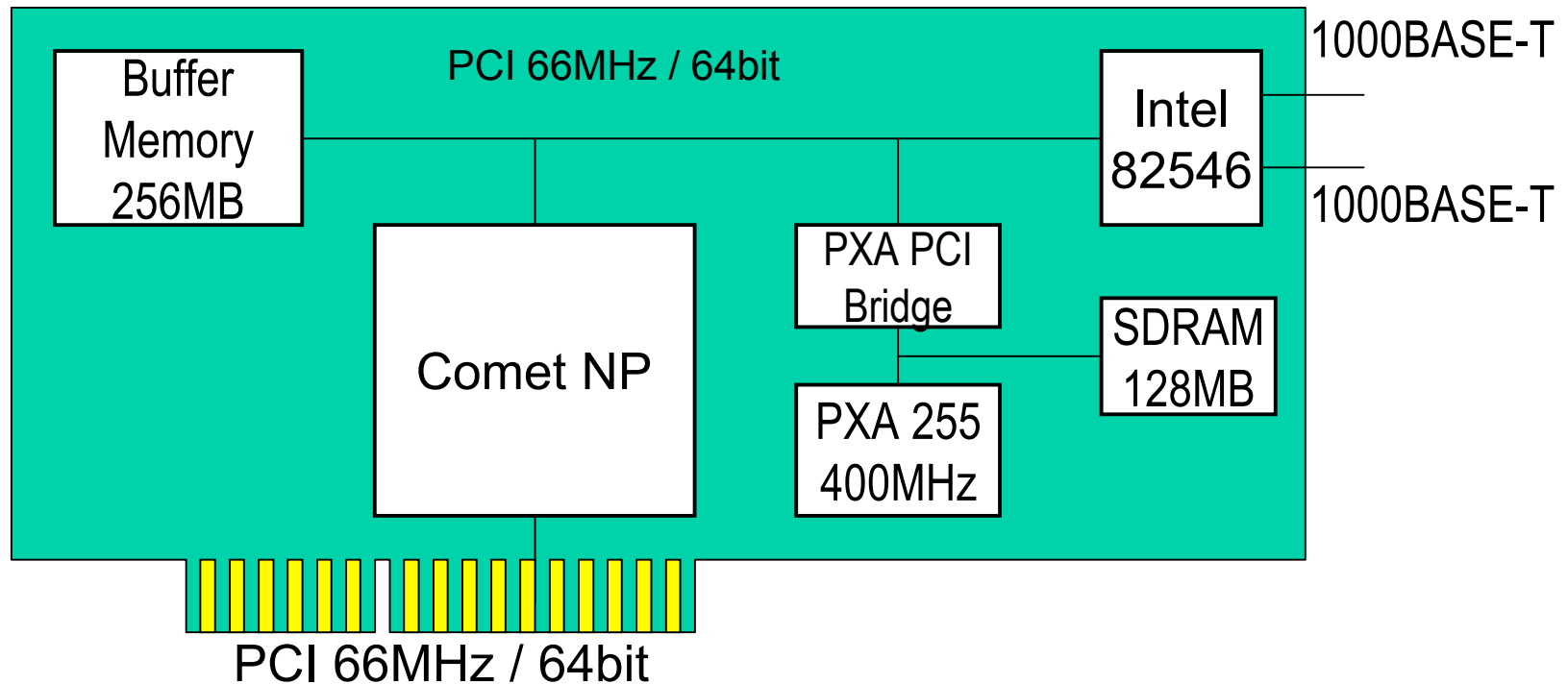


Comet TCP technology

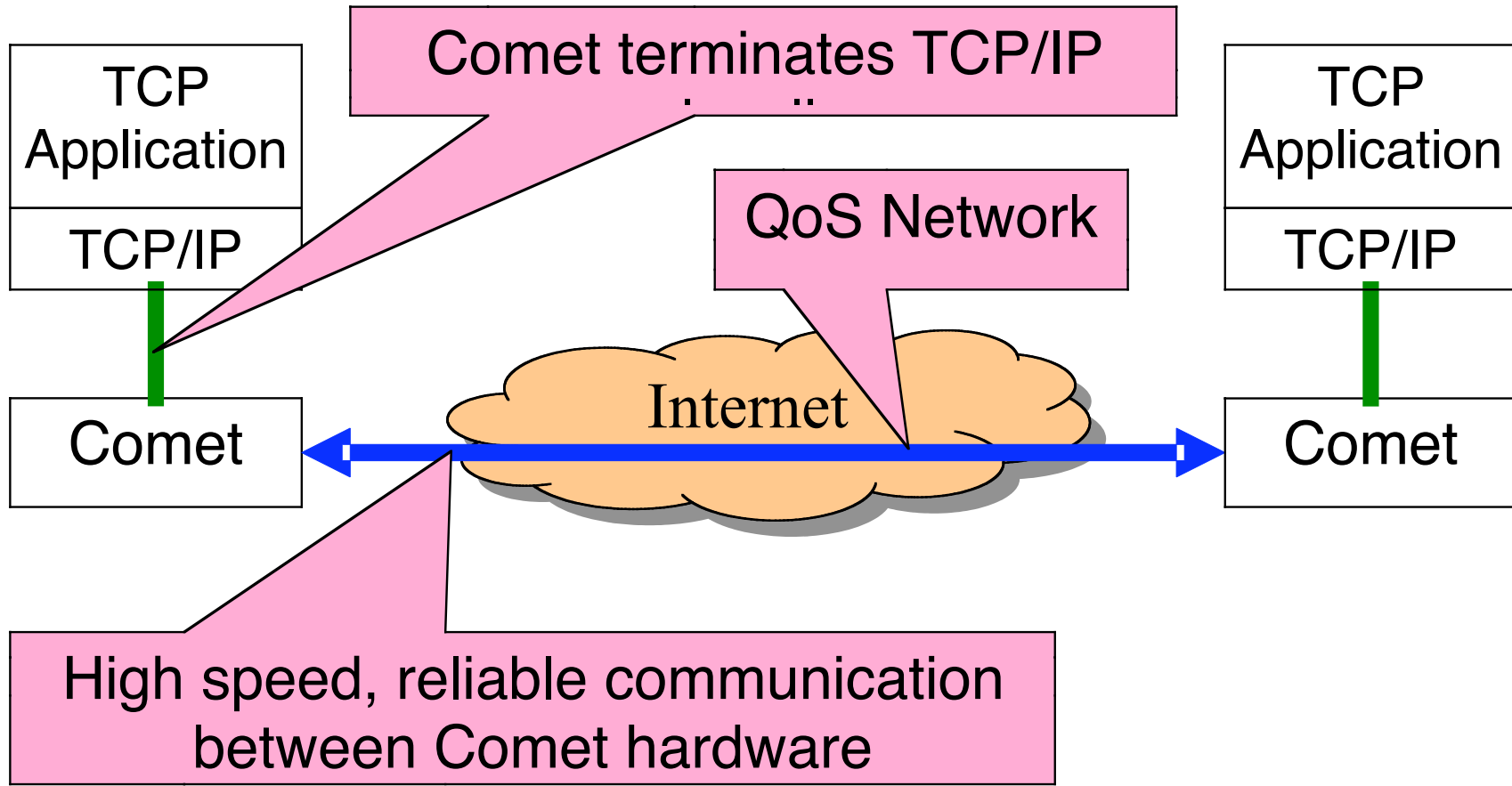
- Low TCP bandwidth due to packet losses
 - TCP congestion window size control
 - Hardware acceleration of TCP by NIC hardware in Long Fat pipe Network (Comet Network Processor)
- Lower CPU overheads for communication
 - Maximum utilization of Network bandwidth (QoS)
 - Out boarding TCP control to NIC
- Encryption / decryption for data security
 - Hardware support for ESP encapsulation
(At BWC, this capability is off)

Comet network processor card

- Normal size PCI NIC card with Comet network processor
- Micro-programmable Comet NP and Xscale CPU

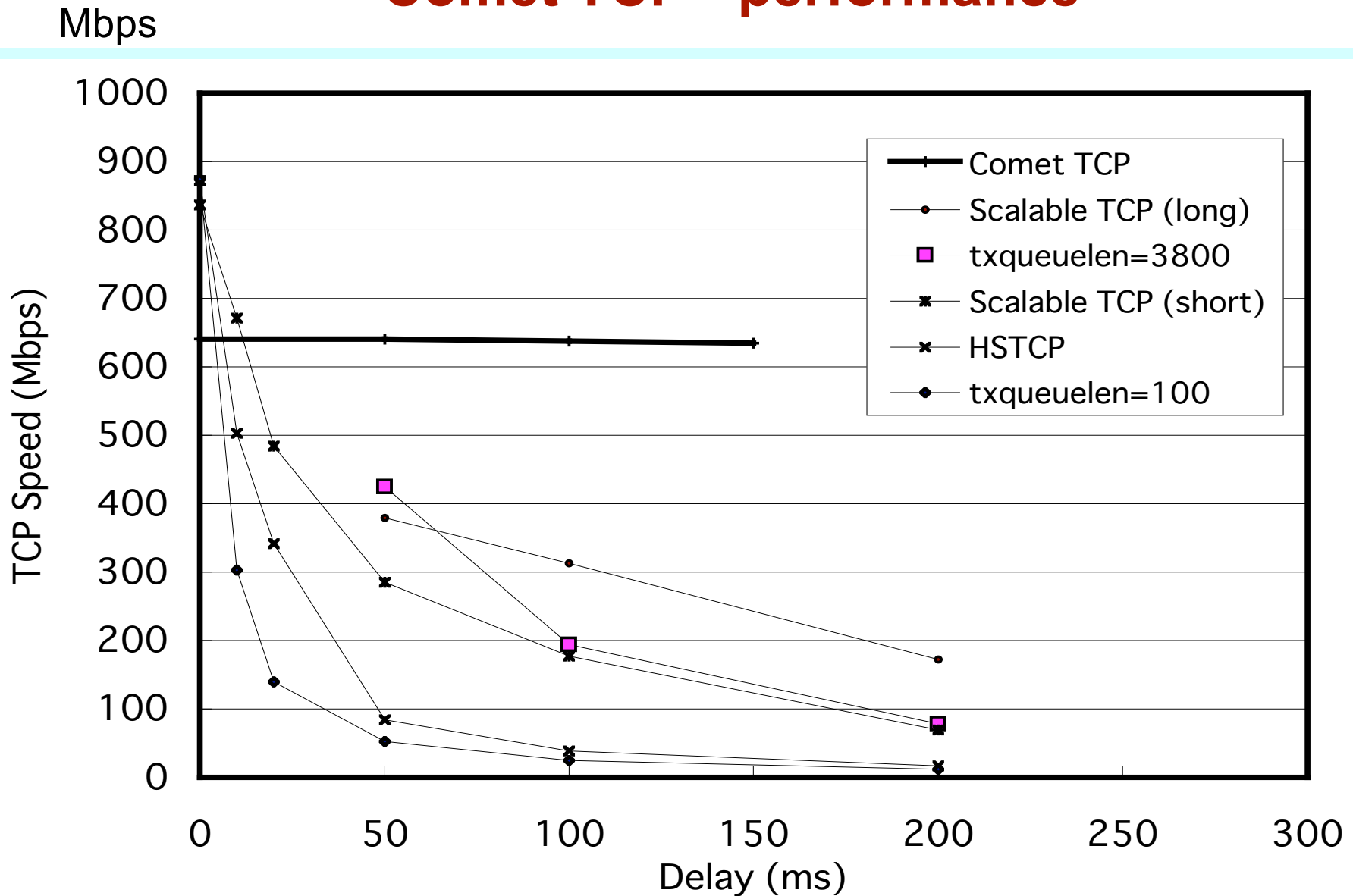


Comet TCP - outline

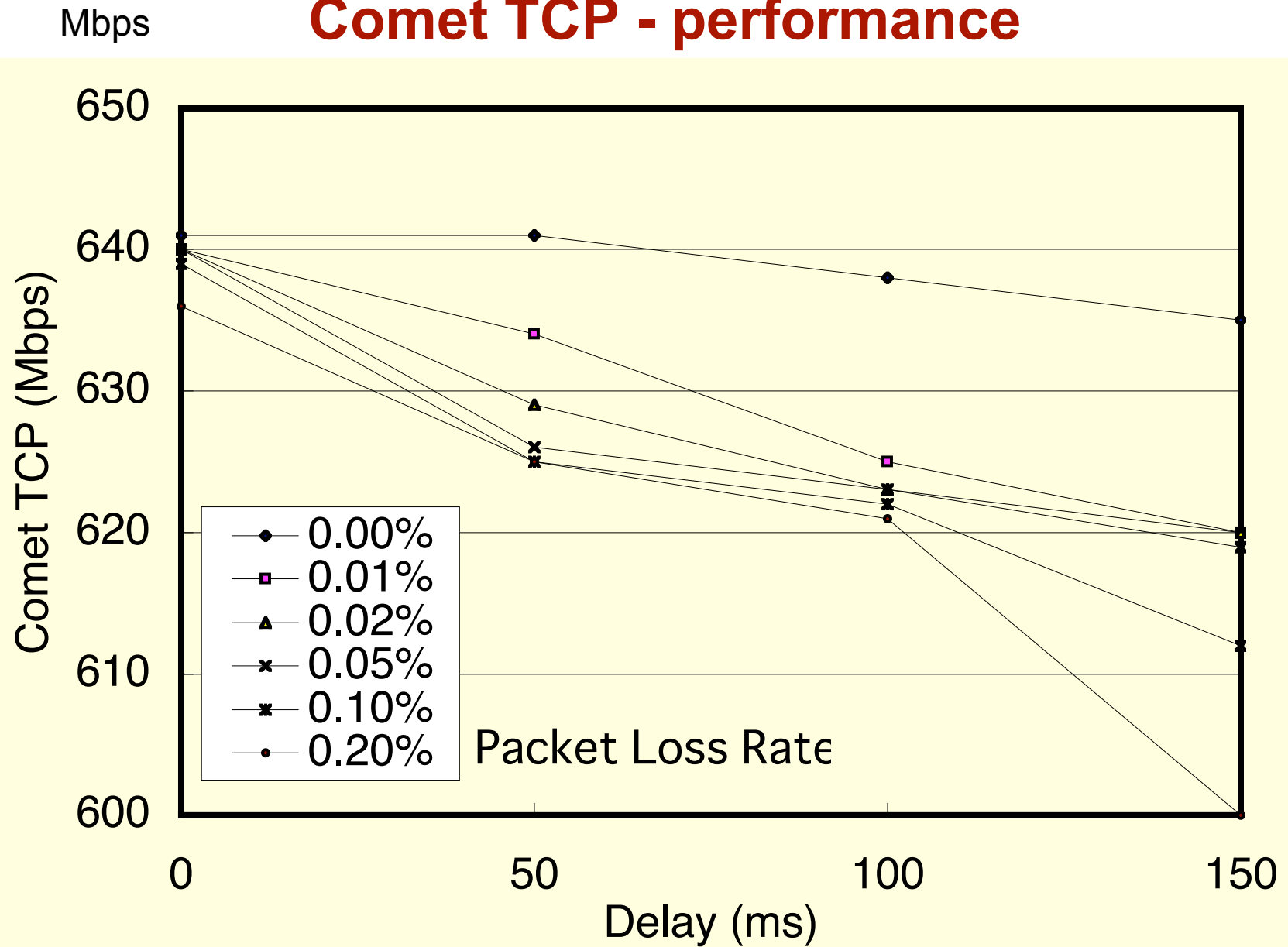


(Comet is already a commercial products)

Comet TCP - performance



Comet TCP - performance



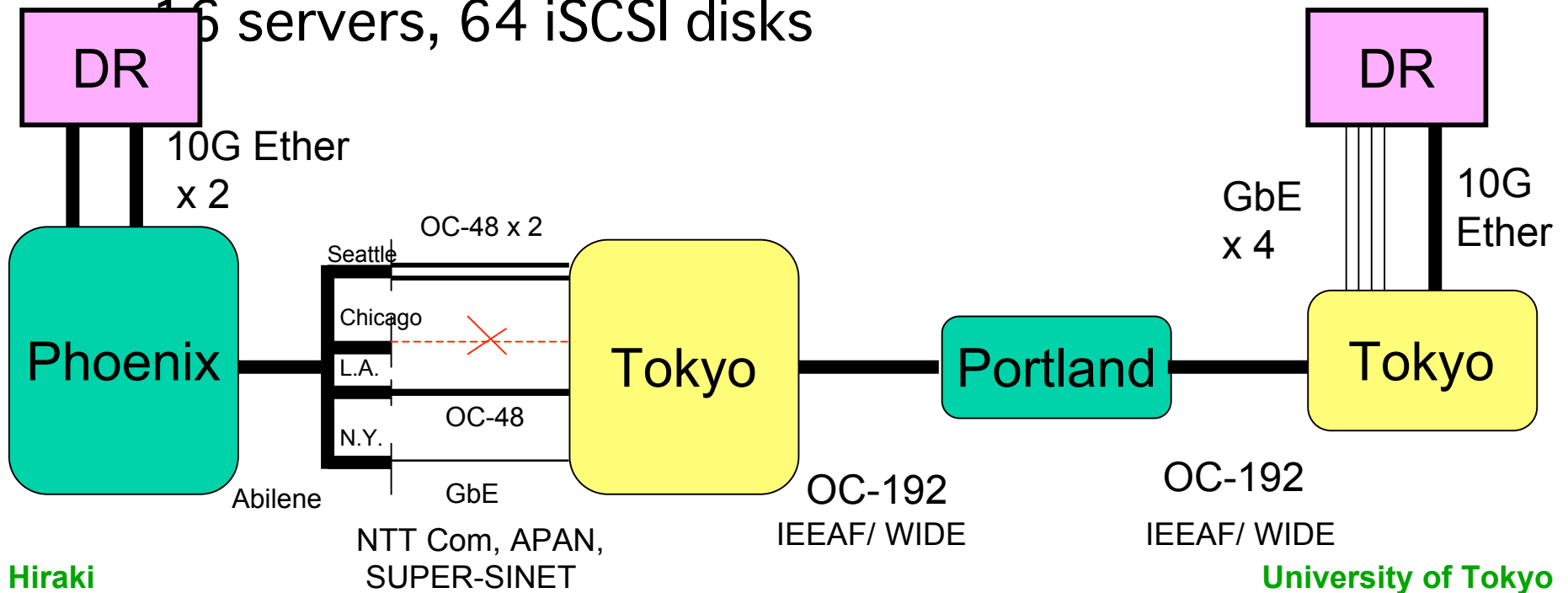
BW2003 US-Japan experiments

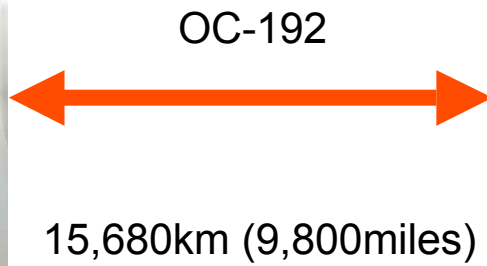
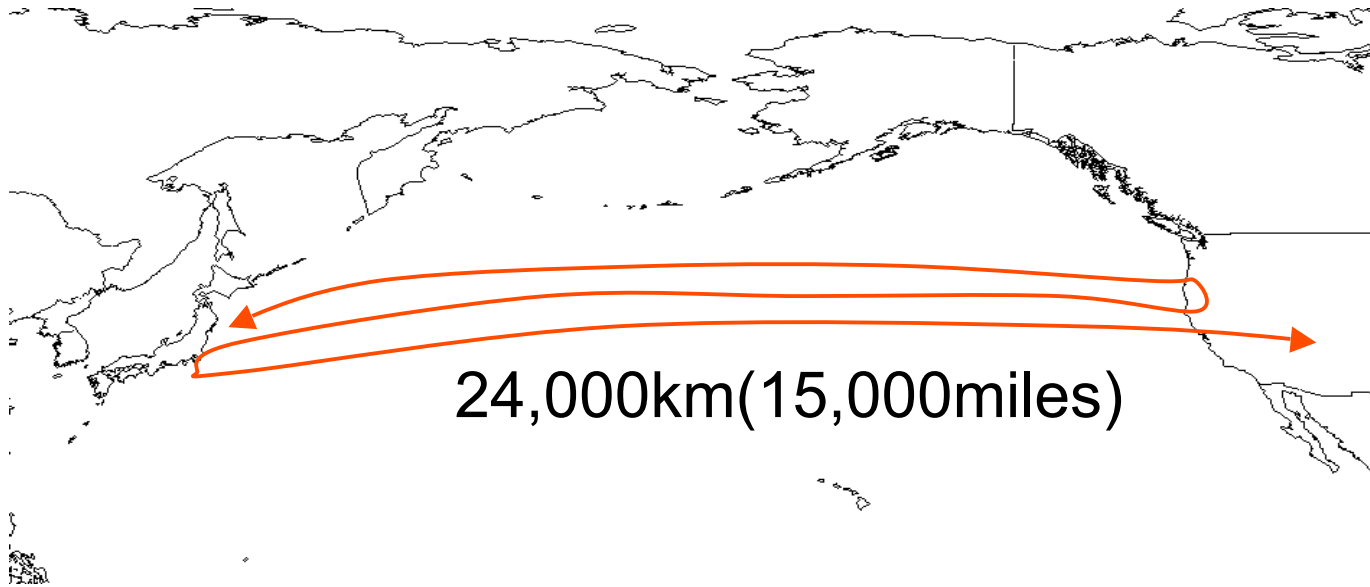
- 24000 km (15,000 miles) distance (~400ms RTT)

_ Phoenix → Tokyo → Portland → Tokyo
 OC-48 x 3 OC-192 OC-192
 GbE x 1

- Transfer ~ 1 TB file

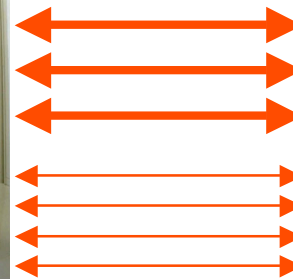
16 servers, 64 iSCSI disks





Juniper
T320

OC-48 x 3
GbE x 4



8,320km
(5,200miles)



Bandwidth during a test run



Results

- Preliminary experiment
 - Tokyo → Portland → Tokyo 15,680km (9,800miles)
 - Peak bandwidth (on network) 8.0 Gbps
 - Average file transfer bandwidth 6.2 Gbps
 - Bandwidth-distance products 125,440 terabit-kilometers/second
- BWC results (pre-test)
 - Phoenix → Tokyo → Portland → Tokyo 24,000 km (15,000 miles)
 - Peak bandwidth (on network) > (8 Gbps)
 - Average file transfer bandwidth > (7 Gbps)
 - Bandwidth-distance products > (168,000 terabit-kilometers/second)
 - More than 10 times improvement from BWC2002 performance

Bad News

- Network cut-down on 11/8
 - _ US-Japan north route connection has been completely out of order
 - _ 2~3 weeks are necessary to repair the under-sea fibers.
 - _ Planned BW = 11.2 Gbps (OC48 x 3 + GbE x 4)
 - Actual maximum BW \doteq 8.2 Gbps (OC48 x 3 + GbE x 1)

How your science benefits from high performance, high bandwidth networking

- Easy and transparent access to remote scientific data
 - Without special programming (normal NFS style accesses)
- Utilization of high-BW network for his data
 - 17 minutes for 1TB file transfer from the opposite location on earth
 - High utilization factor (> 90%)
 - Good for both scientists and network agencies
- Scientists can concentrate on his research topics
 - Good for both Scientists and Computer Scientists

Summary

- **The most distant data transfer** at BWC2003 (24,000 km)
- Hardware acceleration for overcoming latency and decreasing CPU overheads
 - **Comet TCP**, latency tolerant hardware acceleration
 - Same API and interface to user programs
- Possibly **highest bandwidth between Pacific Ocean** for file transfer
- Still **high utilization of available bandwidth**
 - Low level (iSCSI) data transfer architecture

BWC 2003 Experiment is supported by

NTT / VERIO

WIDE
PROJECT

 **Juniper**[®]
NETWORKS

 **APAN**

 **FOUNDRY**[™]
NETWORKS

 **IEEAF**

 **SUPER
SINET**

CISCO SYSTEMS


tyco / Telecommunications