

# Application-Level Measurements Over Dedicated Bandwidth Channels: UltraScienceNet Perspective

Nagi Rao, Bill Wing, Steven Carter, Qishi Wu, Susan Hicks  
Computer Science and Mathematics Division  
Oak Ridge National Laboratory  
raons@ornl.gov

<https://www.usn.ornl.gov>

Circuit and Lightpath Measurement Panel  
Winter 2006 ESCC/Internet2 Joint Techs Workshop  
February 7, 2006, Albuquerque, NM

Research Sponsored by  
High-Performance Networks Program

OAK RIDGE NATIONAL LABORATORY  
U. S. DEPARTMENT OF ENERGY

U.S. Department of Energy

  
UT-BATTELLE

# Contents

- **My Background**
- **USN Overview**
- **A-A Performance Profiles**
  - **UDP Throughput Profiles**
  - **Transport Protocol Design**
- **End-to-End Measurements Wish List**

# My Background

- **Co-PI of DOE UltraScience Net Project along with Bill Wing (ORNL)**
- **Co-PI NSF CHEETAH Project along with Malathi Veeraraghavan (UVA), Ibrahim Habib (CUNY), John Blondin (NSSU)**
  
- **Distinguished R&D Staff, been at ORNL since 1993**
- **PhD in Computer Science (1988)**
- **Network Researcher:**
  - **Advanced Bandwidth Scheduling**
  - **Transport Protocols**
    - **AIMD Dynamics – Chaos; Stabilized flows; High-utilization of dedicated channels**
- **Other Areas: sensor-cyber networks, sensor fusion, robot navigation**

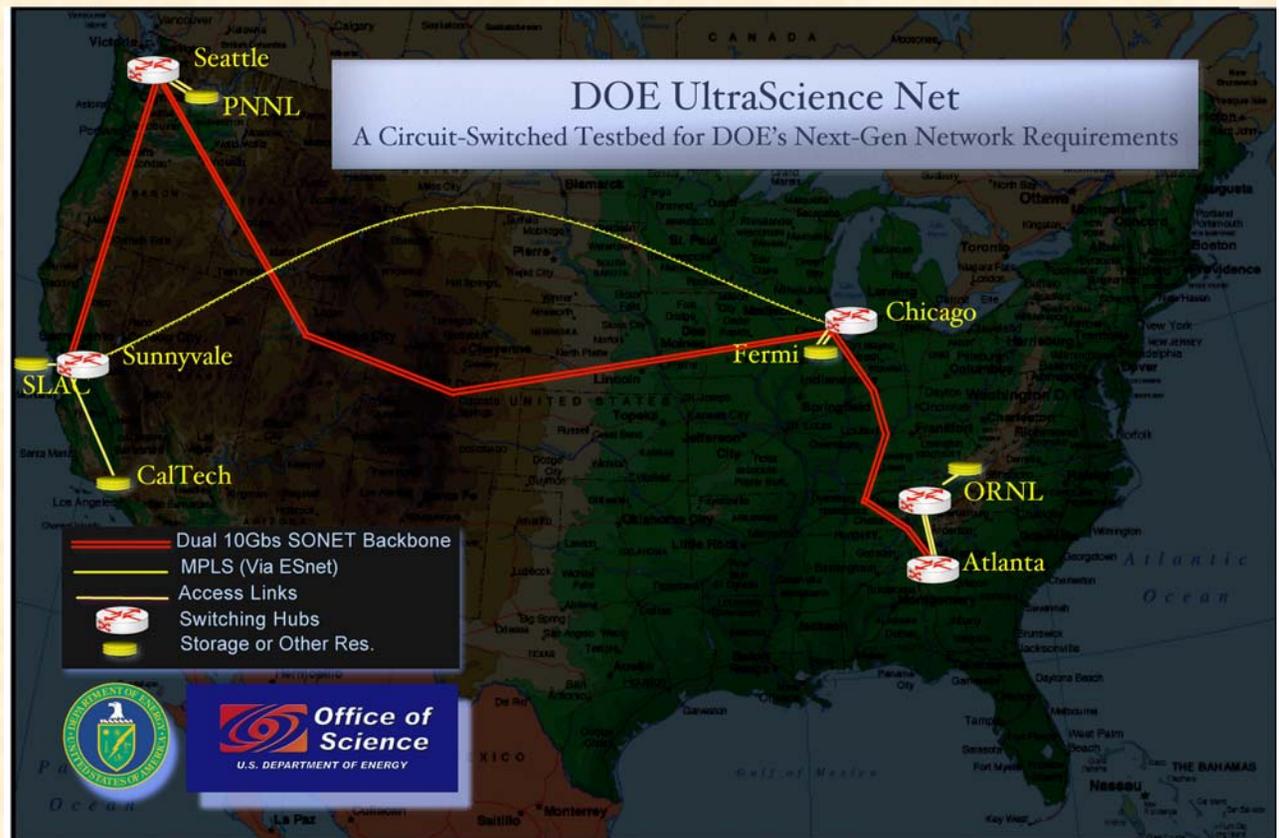
# DOE UltraScience Net – In a Nutshell

## Experimental Network Research Testbed:

To support advanced networking and related application technologies for DOE large-scale science projects

### Features

- End-to-end guaranteed bandwidth channels
- Dynamic, in-advance, reservation and provisioning of fractional/full lambdas
- Secure control-plane for signaling
- Proximity to DOE sites: NLCF, FNL, NERSC
- Peering with ESnet, NSF CHEETAH and other networks



# DOE UltraScience Net: Need, Concept and Challenges

## The Need

- DOE large-scale science applications on supercomputers and experimental facilities require high-performance networking
  - Moving petabyte data sets, collaborative visualization and computational steering (all in an environment requiring improved security)
- Application areas span the disciplinary spectrum: high energy physics, climate, astrophysics, fusion energy, genomics, and others

## Promising Solution

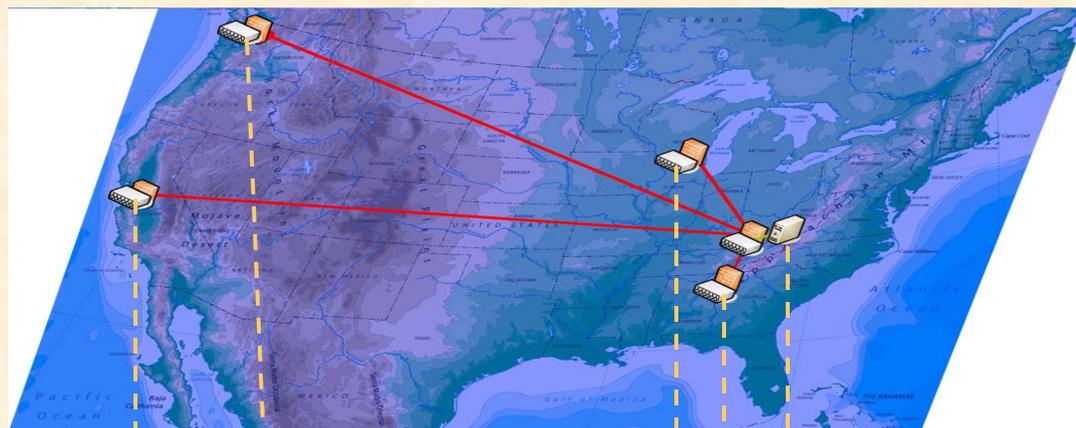
- High bandwidth and agile network capable of providing scheduled dedicated channels: multiple 10Gbps to 150 Mbps
- Protocols are simpler for high throughput and control channels

## Challenges: Several technologies need to be (fully) developed

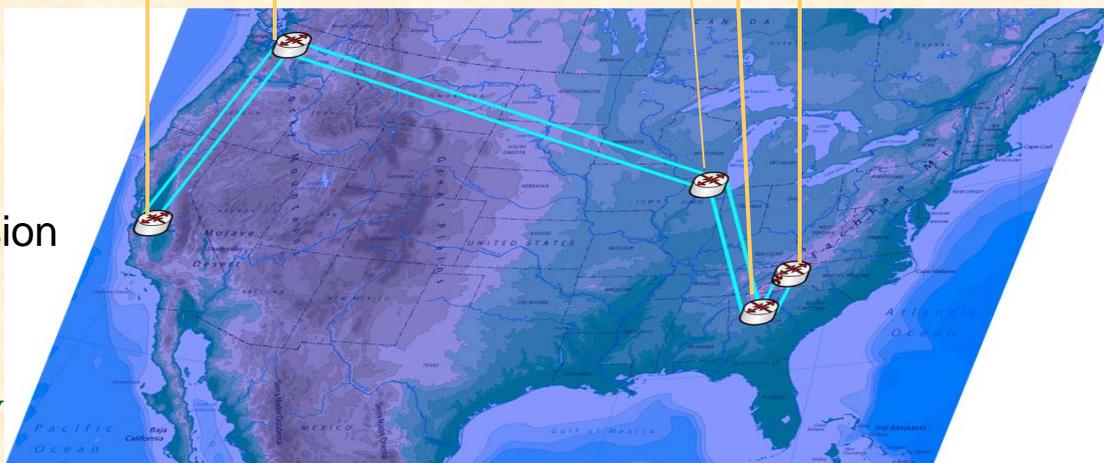
- User-/application-driven agile control plane:
  - Dynamic scheduling and provisioning
  - Security – encryption, authentication, authorization
- Protocols, middleware, and applications optimized for dedicated channels

# USN Architecture: Separate Data-Plane and Control-Planes

Secure control-plane with:  
Encryption, authentication and  
authorization  
On-demand and advanced  
provisioning



Dual OC192 backbone:  
SONET-switched in the  
backbone  
Ethernet-SONET conversion



**OAK RIDGE NATIONAL LABORATORY**  
**U. S. DEPARTMENT OF ENERGY**



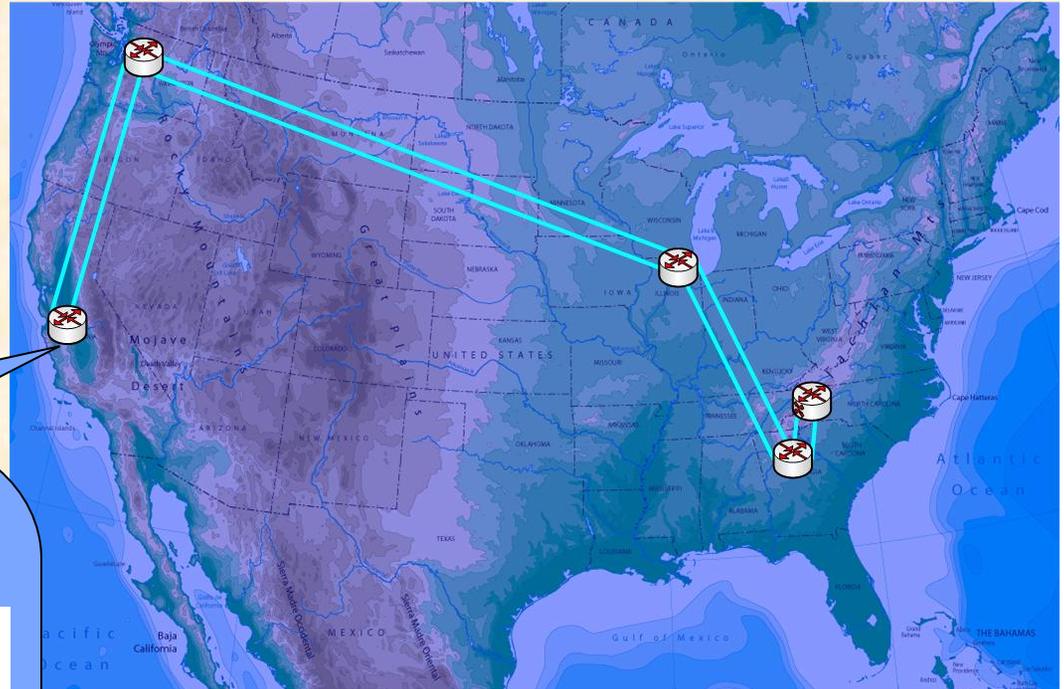
# USN Data-Plane: Node Configuration

## In the Core:

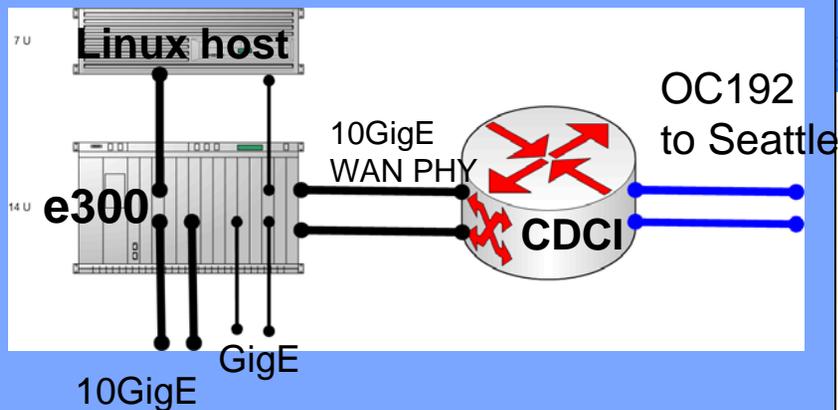
- Two OC192 switched by Ciena CDCIs

## At the Edge

- 10/1 GigE provisioning using Force10 E300s



## Node Configuration



Connections to  
CalTech and ESnet

## Data Plane User Connections:

Direct connections to:

core switches –SONET &1GigE

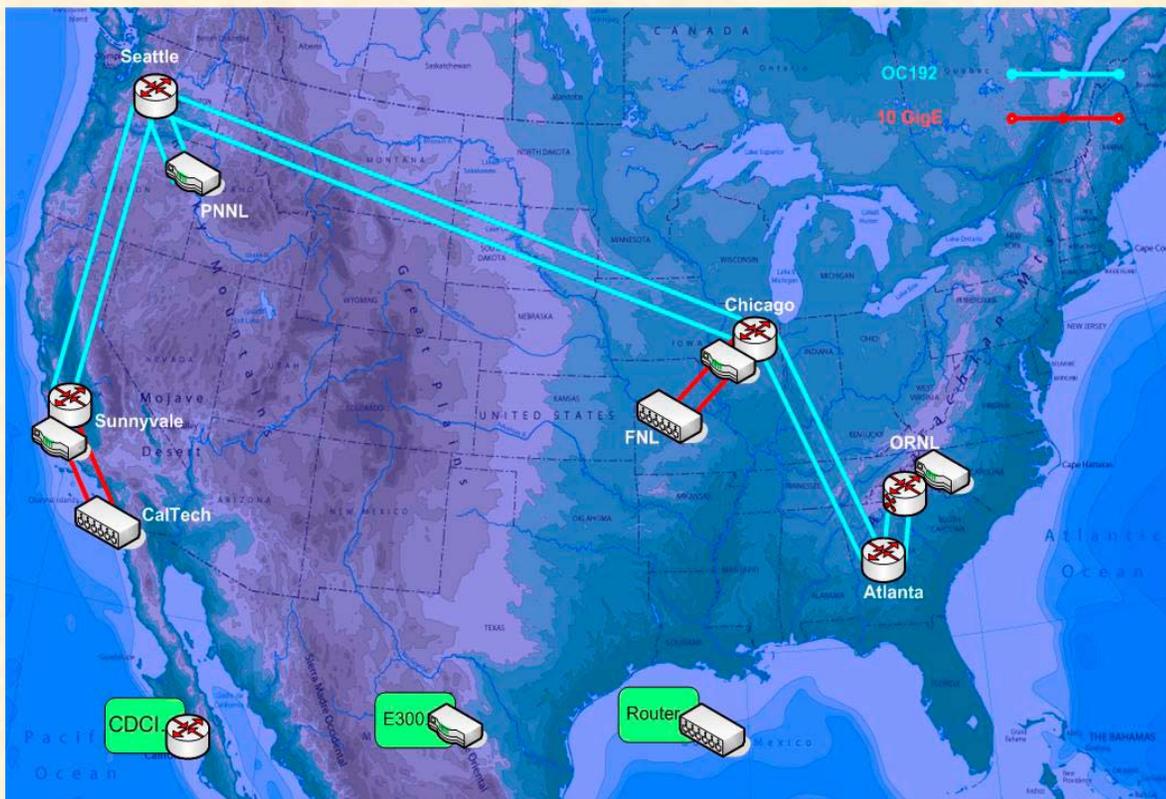
MSP – Ethernet channels

Utilize UltraScience Net hosts

# DOE UltraScience Net: Monitoring

**Core Switches: CDCI – Ciena Node Manager**

**Edge Switches: E300 – Cacti using ICMP (packet counts)**



**OAK RIDGE NATIONAL LABORATORY**  
**U. S. DEPARTMENT OF ENERGY**

**UT-BATTELLE**

# End-to-End Measurements on USN

## Motivation:

**USN is built to provide dedicated bandwidth channels to Applications:**

- **What type of throughput is seen at applications?**
  - **Reasonable Expectation: Throughput is**
    - close to channel bandwidth and
    - stable if rate-controlled transport is used
  - **Measurements indicate: throughput is not always the channel bandwidth nor has stable dynamics!**

## Needs Specific to USN:

**End-to-End Application Throughput (EEAT):**

**For high-bandwidth applications: Channel utilization**

**For control-application: Transport dynamics**

**Yes, outside the domain traditional networking problem space**

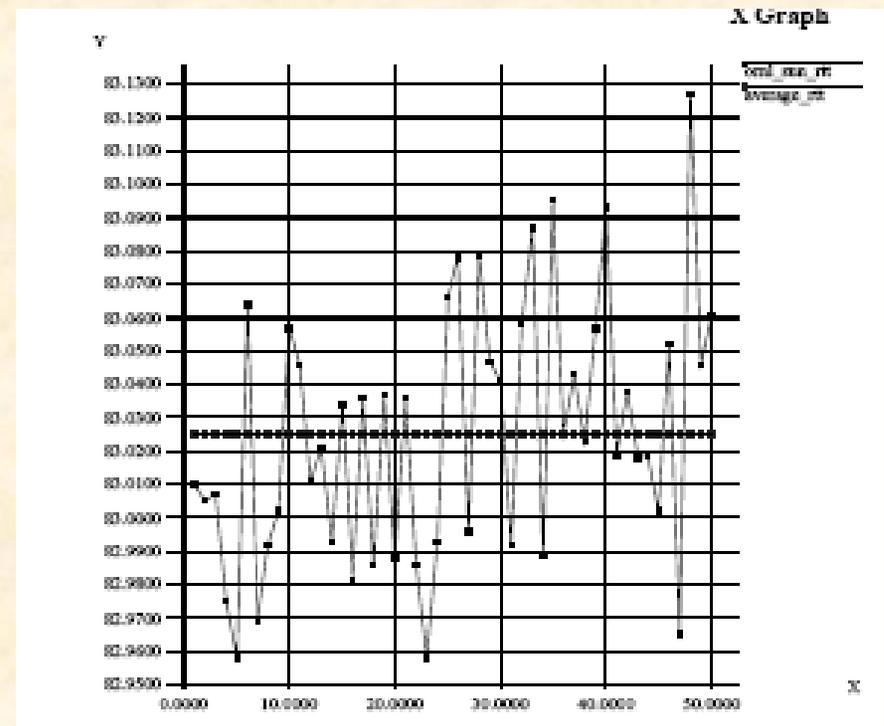


# Some Experimental Results

- Layer-2 double-loopback test:
  - Entire USN SONET backbone connected in 16000 mile single connection
  - 16 hours continuous zero SONET-level errors

- **Jitter measurements**

- **ORNL-SUNNYVALE host-to-host 1K packets**
- **round-trip time:**
  - mean: 82ms
  - jitter: 0.2%



# Throughput Profile

## Plot of receiving rate as a function of sending rate

Its precise interpretation depends on:

- Sending and receiving mechanisms
- Definition of rates

For protocol optimizations, it is important to use its own sending mechanism to generate the profile

### Window-based sending process for UDP datagrams:

Send  $W_c(t)$  datagrams in a one step – *window size*

Wait for  $T_s(t)$  time called *idle-time* or *wait-time*

Sending rate at time resolution  $T_s(t)$ :

$$r_s(t) = \frac{W_c(t)}{T_s(t) + T_c(t)}$$

This is an adhoc mechanism facilitated by 1GigE NIC

# Throughput Profile: Internet Connection - ORNL-LSU

Throughput and loss rates vs. sending rate (window size, cycle time)

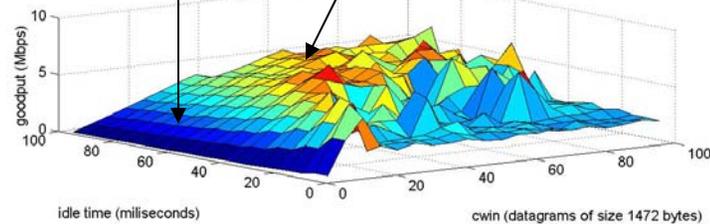
Typical day

Christmas day

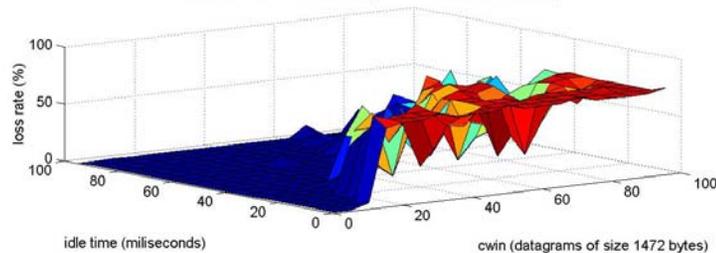
Stabilization zone

Goodput vs. cwin & idle time (Mon Dec 02 20:37:04 2002)

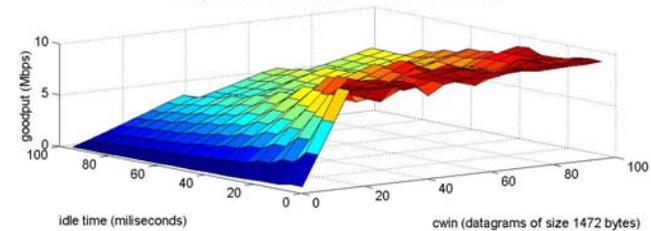
Peak zone



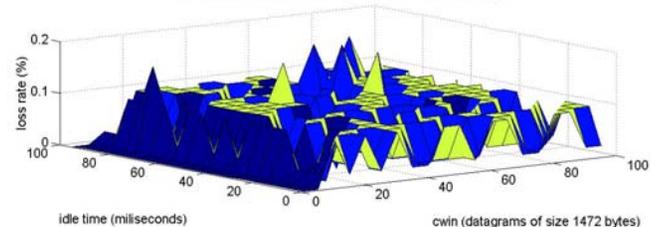
Loss rate vs. cwin & idle time (Mon Dec 02 20:37:04 2002)



Goodput vs. cwin & idle time (Wed Dec 25 18:02:04 2002)

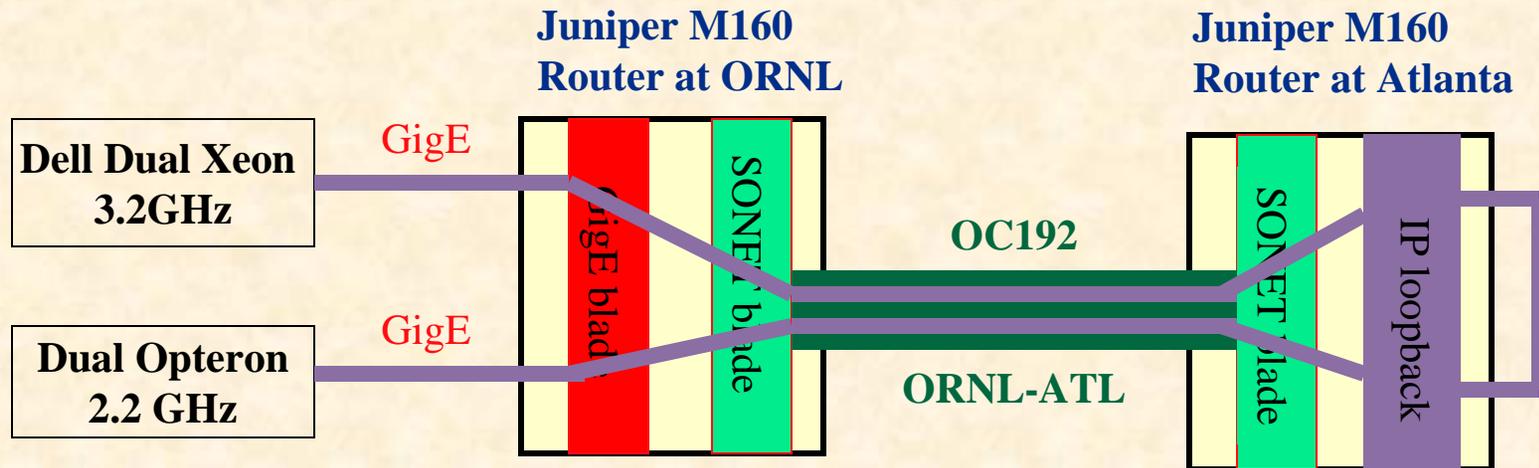


Loss rate vs. cwin & idle time (Wed Dec 25 18:02:04 2002)



Objective: adjust source rate to yield the desired throughput at destination

# 1Gbps ORNL-ATL-ORNL Dedicated IP Channel

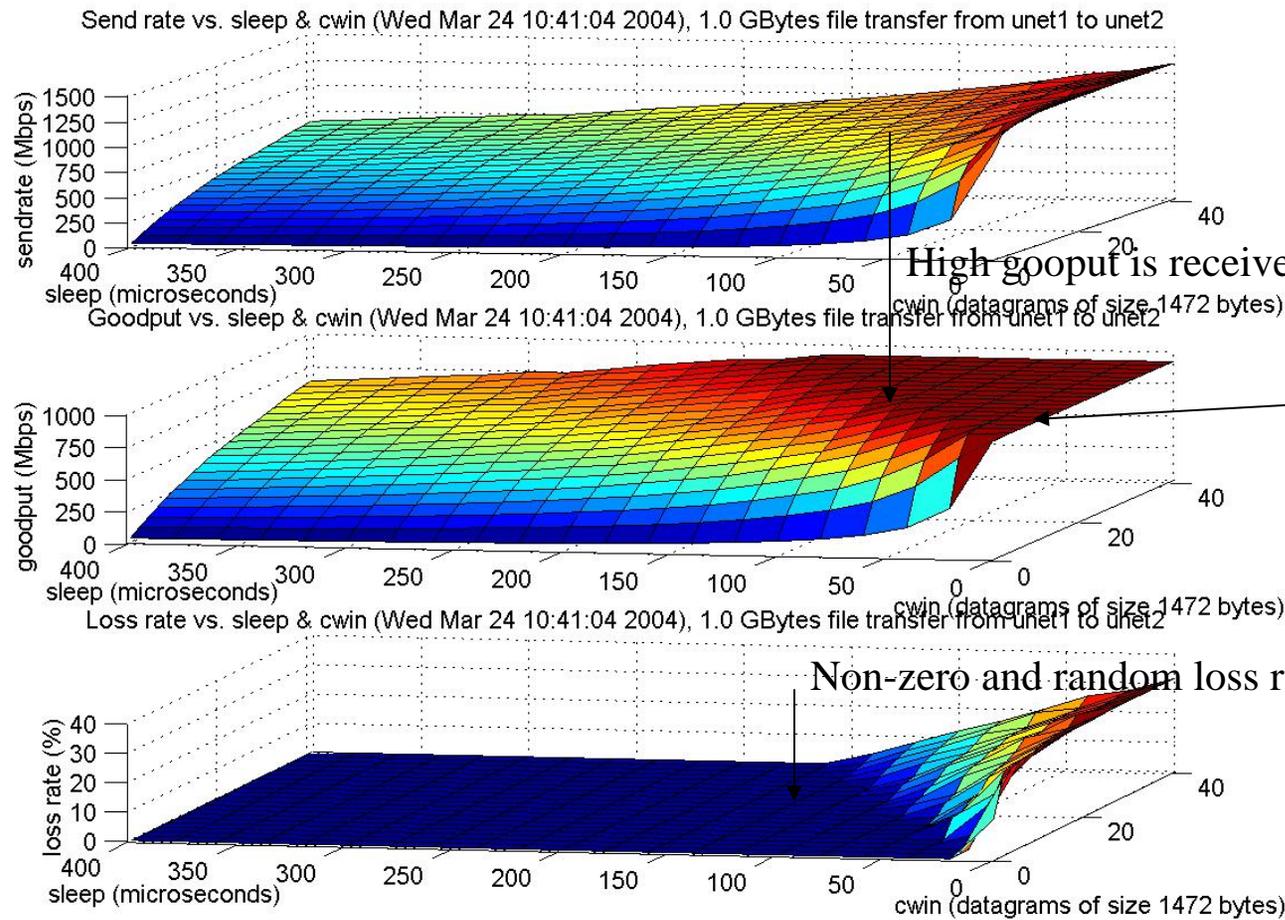


- **Non-Uniform Physical Channel:**
  - GigE - SONET - GigE
  - ~500 network miles
- **End-to-End IP Path**
  - Both GigE links are dedicated to the channel
  - Other host traffic is handled through second NIC
- **Routers, OC192 and hosts are lightly loaded**
- **IP-based Applications and Protocols are readily executed**

# Dedicated Hosts

- **Hosts:**
  - **Linux 2.4 kernel (Redhat, Suse)**
  - **1/10GigE NICS:**
    - **optical connection to Juniper M160 or Force10 E300**
    - **copper connection Ethernet switch/router**
  - **Disks: RAID 0 dual disks (140GB SCSI)**
  - **XFS file system**
    - **Peak disk data rate is ~1.2Gbps (IO Zone measurements)**
    - **Disk is not a bottleneck for 1Gbps data rates**

# UDP goodput and loss profile



High gooput is received at non-trivial loss

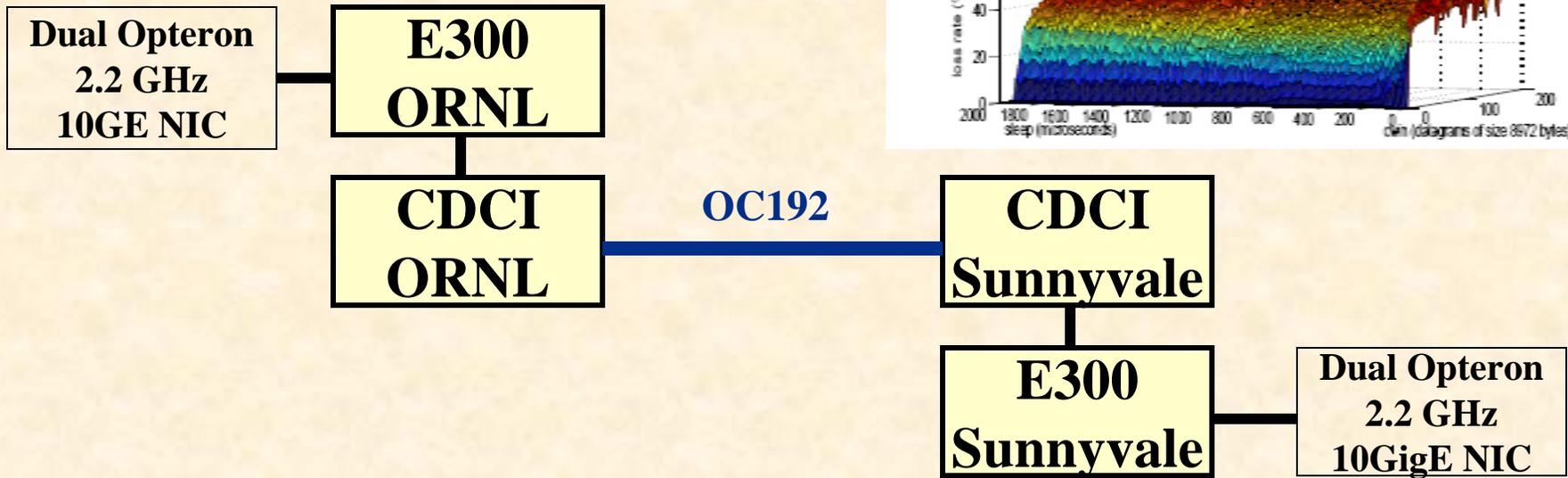
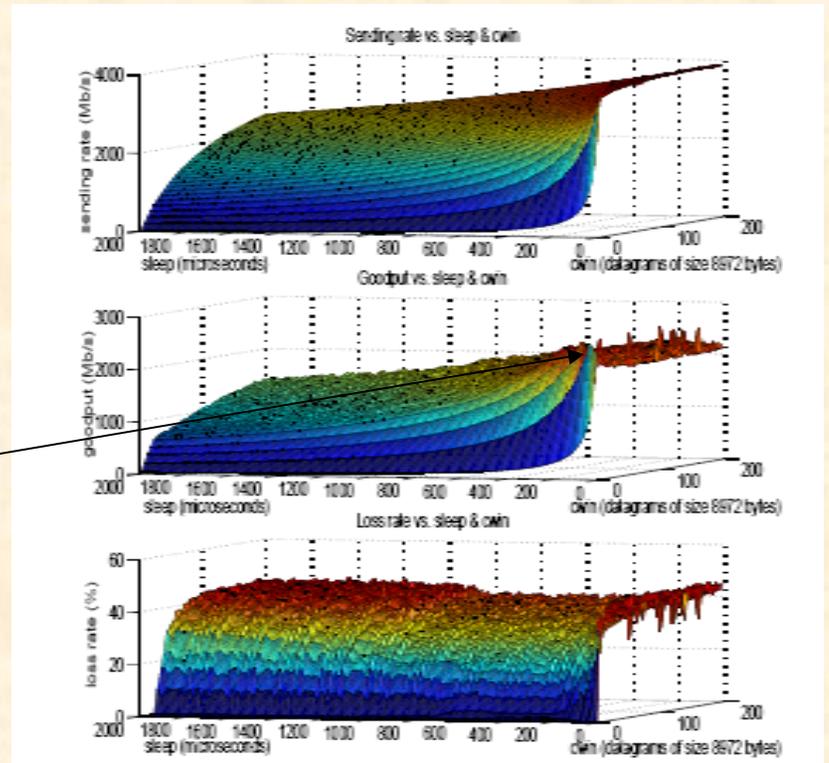
Gooput plateau ~990Mbps

Non-zero and random loss rate

Point in horizontal plane:  $(W_c(t), T_s(t))$

# Throughput profile USN ORNL-SUNNYVALE

- Transport measurements between linux hosts with 10GigE NICS
  - ORNL-SUN host-to-host file transfers 4000mile, 10G connection
  - Limited by host - Hurricane
  - Average throughput 2.3Gbps
  - Loss rate < 0.1%



# Channel Profiles

- **Throughput Profiles**
  - Provide valuable EEAT information – UDP-based transport
    - Peak achievable throughput
    - Dynamics of throughput
  - Transport Protocols can be optimized:
    - Need to stabilize the operating point
      - HURRICANE Protocol – not flow-friendly; peak utilization
      - RUNAT – stochastic maximization of throughput

**We need comprehensive channel profiling capability:**

- What class of channel profiles are appropriate?
- How to measure and present them?

# Integrated On-line Channel Profile Capability :

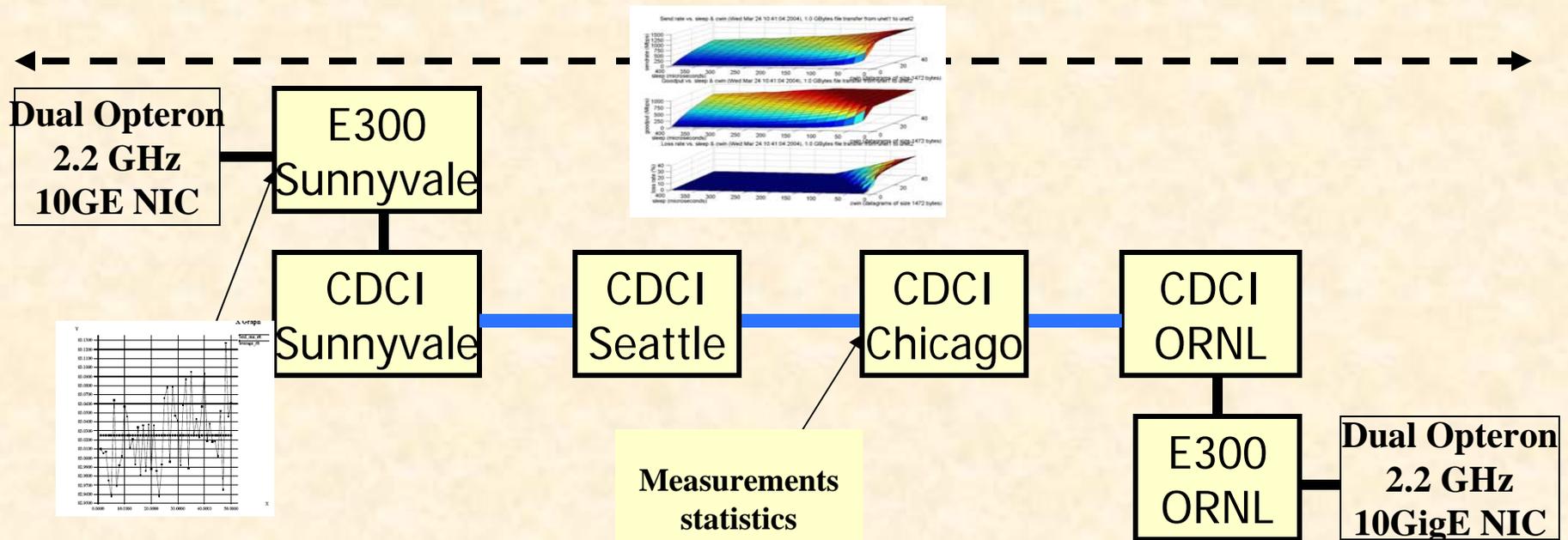
On-line throughput profiles with detailed decomposition

Graphical high-level profile  
connection overlaid on map

Annotated statistics and measurements:  
connection, link, NIC, host, application

Ability to bring-up individual components  
profiles statistics, etc

Tightly-coupled analysis and diagnosis tools



## **Integrated Capability for On-line Channel Profiles**

### **Complex Task:**

**Combine various measurements:**

**connection, link, port, NIC, host**

**Intelligent fusion of all information:**

**multiple-level analysis and diagnosis**

**Level of intrusiveness**

**Support for active diagnosis**

**Assembling multiple tools**

# **Conclusions**

**Measurements for dedicated channels is a new frontier:**

- We are beginning to understand the needs**
  - Applications and transport play an integral role**
- Very complex task: needs efforts from multiple domains**
  - Need beyond traditional layer-3 tools**
    - Some layer-2 connections may carry non-IP traffic, eg FiberChannel**

**Thank you**  
**<https://www.usn.ornl.gov>**

**OAK RIDGE NATIONAL LABORATORY**  
**U. S. DEPARTMENT OF ENERGY**

