

# Towards Usable Terabit WAN's: The OptIPuter System Software



Andrew A. Chien  
Director, Center for Networked Systems  
SAIC Chair Professor, Computer Science and Engineering  
University of California, San Diego

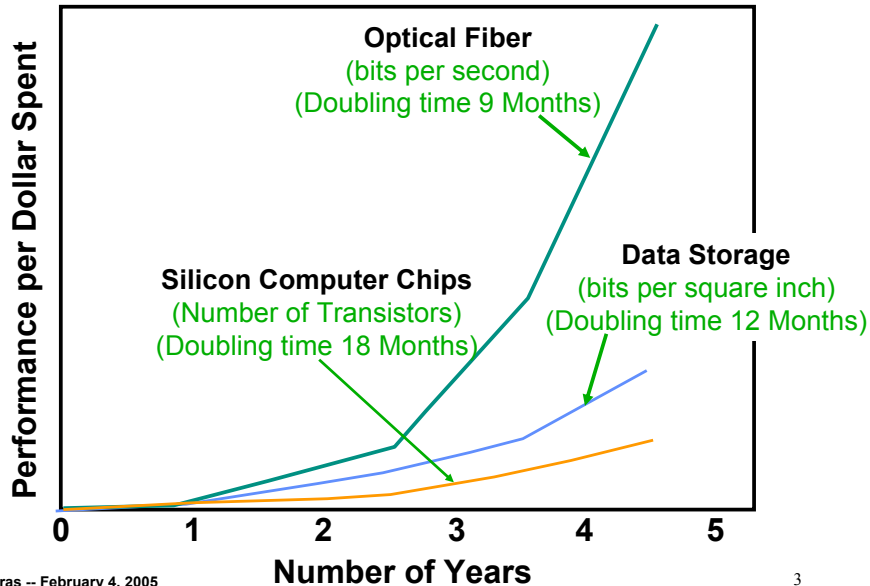
Mardi Gras Conference  
Center for Computation and Technology  
Louisiana State University  
February 4, 2005



## Outline

- **The Opportunity: Lambda-Grids**
  - Applications Drivers and Testbeds
- **OptIPuter System Software**
  - Model of Use: Distributed Virtual Computers
  - High Performance Communication
  - Supporting Data-Intensive Applications
- **Demonstrations**
  - Terabit Juggling
  - 3-Layer OptIPuter DVC Demonstration
- **Related Work**
- **Summary and Future Work**

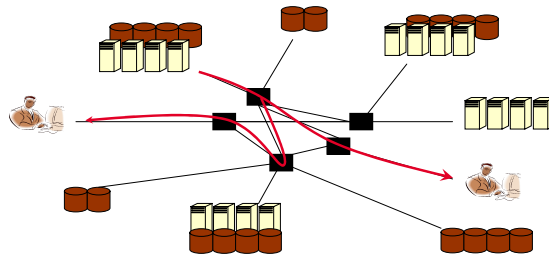
# Optical Networks Are Emerging as the 21<sup>st</sup> Century Driver



3

# Lambda Grids Empower Distributed Resource Sharing and Collaboration

- **On-Demand End-to-End Optical Connections**
  - Dedicated High Bandwidth: Close Coupling
- **Grids**
  - Flexible, Open Resource Sharing
- **Lambda Grids = Grids Powered by Dedicated Lambdas**
  - Dynamically Constructed, Distributed Resource Collections
  - Communicating through Dedicated Connections

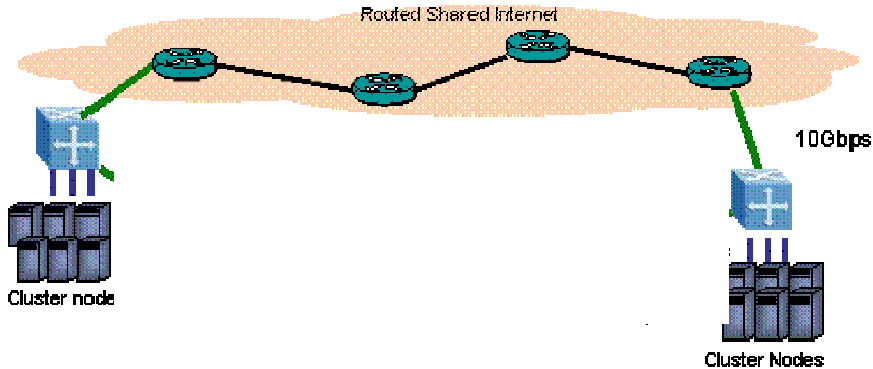


Mardi Gras -- February 4, 2005

4

# OptIPuter Physical Network

- **Grids -- Shared Internet on underlying Physical Telecom Infrastructure**
- **Lambda Grids -- Dedicated Optical Paths on underlying Physical Telecom Infrastructure**
- **New End-to-End Capabilities: Extraordinary Bandwidth, Private Connections**



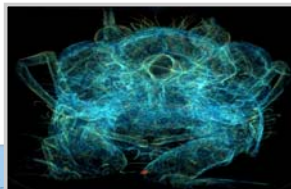
5

## OptIPuter Project: Explore Impact of Lambdas

- **OptIPuter: System Software for Lambda Grids for next-generation E-science**
  - International Testbed for Experimentation (UCSD, UIC, UCI, Amsterdam, etc.)
  - Leading E-science Drivers (Neuroscience, Geophysical/Earth Sciences)
    - 3-D Data Analysis, Visualization and Collaboration Applications
    - Data-intensive and Real-time, Distributed data sources/sinks
  - Wealth of Innovative System Software Research (protocols, DVC, storage, etc.)



Mardi Gras -- Febru



**Smarr, Papadopoulos,  
Ellisman, Orcutt,  
Chien – UCSD  
DeFanti, Leigh - UIC**

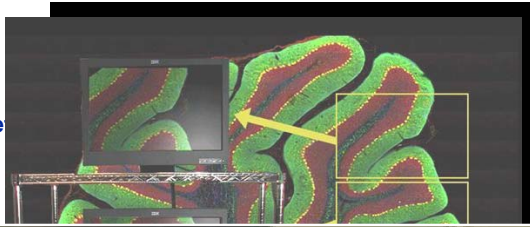


<http://www.optiputer.net>

# Large-Scale, Data-intensive Application Drivers

## 1. Geoscience Imaging

- 10B Pixel Images, PB Database



## 2. Neuroscience Imaging

- 1TB Images, 1 PB



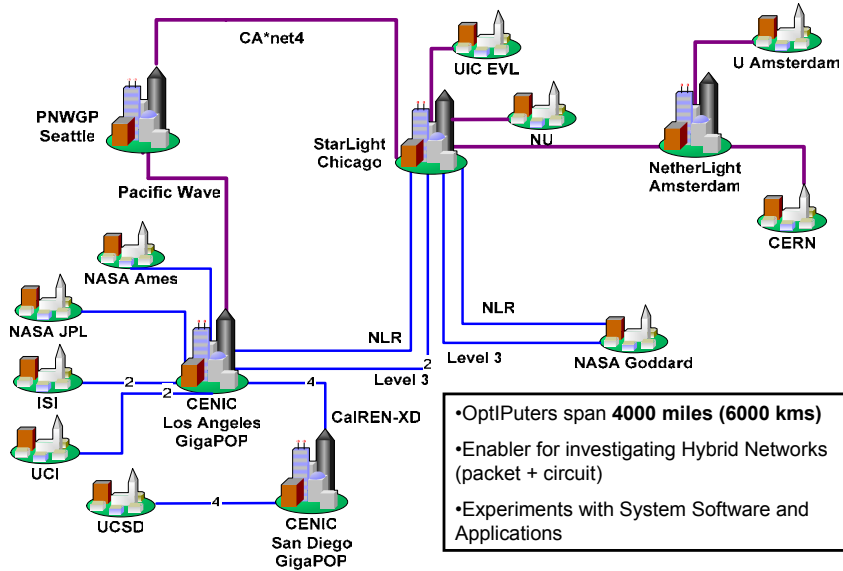
## 3. Animal Observation

- 10GB days/room, 100TB/year



# Large-Scale Testbeds

# The OptIPuter LambdaGrid



Mardi Gras -- February 4, 2005

[DeFanti & Papadopoulos, UCSD]

10

# OptIPuter System Software Challenges

- **What is the Model of Use for Dynamic Lambdas?**
- **How Do We Exploit the Communication Capabilities of Lambdas?**
- **How Do We Support Emerging Data-intensive Applications?**

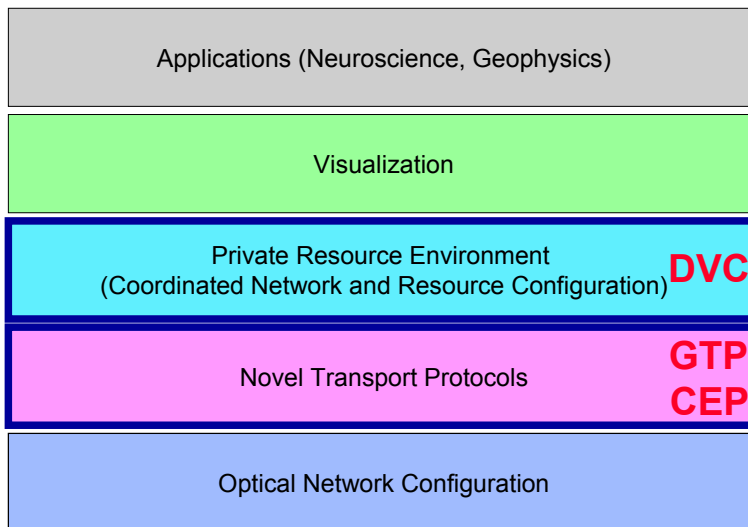
Mardi Gras -- February 4, 2005

11

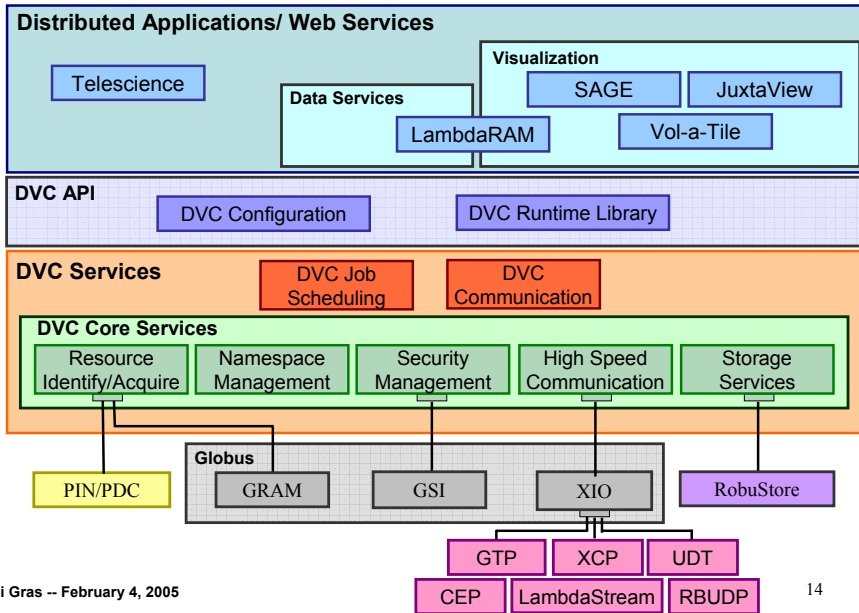
# Models of Use

- **1. Automatic Flow Optimization**
  - End Systems or Network Detects Large *Flows* and Configures Optical Paths to Optimize Extant Flows
  - Intelligent Network, Optimizes for Application and Network Flow Performance [various, BigBangwidth]
- **2. Scheduled Transfers: Optimized FTP [Cheetah, Veeraraghavan03]**
  - Applications Request File *Transfers*
  - Network Schedules and Configures Dedicated Paths
  - Optimizes Network and End Systems for File Transfers
- **3. Distributed Virtual Computer: Grid with High Performance Private Network [DVC, Taesombut&Chien04]**
  - System View of a Grid Resource Collection
  - Private *Network* Constructed and Managed for High Performance
  - Lambda Grid: Dedicated Lambda's + Grid Resource Collection
  - Integrates Resources, Networks in SAN-like Fashion

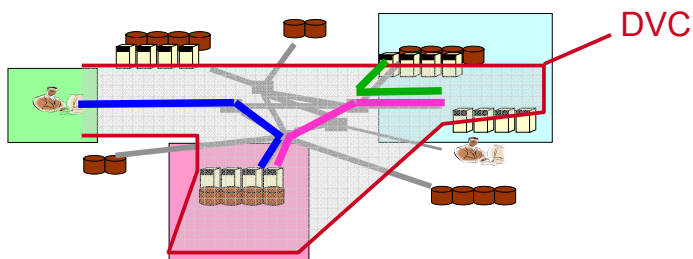
# OptIPuter Software "Stack"



# OptIPuter Software Architecture v1.5

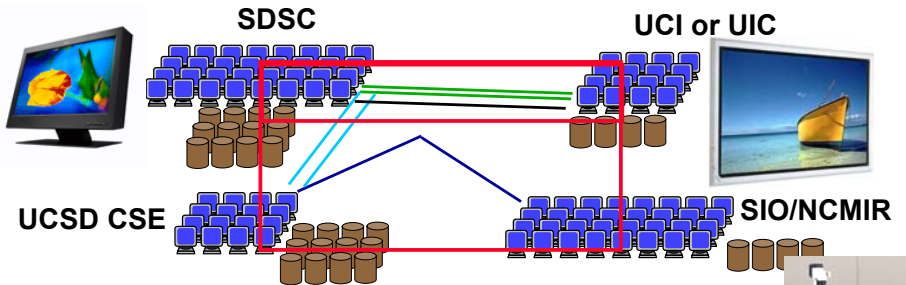


# Distributed Virtual Computer (DVC)



- **Application Requests Grid Resources AND Network Connectivity**
  - Redline-style Specification, 1<sup>st</sup> Order Constraint Language
- **DVC Broker Establishes DVC**
  - Configures Lambdas, Network, Binds Ends Resources
  - Leverages Grid Protocols for Security, Resource Access
  - DVC <-> Private Resource Environment, Surface thru WSRF
- **Key Advantages**
  - Simple Application of Complex Network/Resource Environment
  - Single Interface to Novel Communication Primitives/Protocols

# DVC Examples



- **TeleMicroscopy Experiment DVC**
  - Microscope + Compute Resources + Storage System
  - Dedicated Lambdas + Switching
- **Collaborative Visualization Real-Time DVC**
  - Grid Resources + Real-Time (TMO, Kim, UCIrvine)
  - Dedicated Lambdas + Switching
  - Photonic Multicast or LambdaRAM (Leigh, UIChicago)

Mardi Gras -- February 4, 2005

16

## EVL's JuxtaView: Viewing Extremely High-Resolution Data on the GeoWall<sup>2</sup>

- Data sets have a real need for display resolution.
- JuxtaView copies data across all cluster nodes as memory-mapped files.
- Next phase is to use LambdaRAM for remote memory access.
- Need to examine JuxtaView's memory access patterns to provide heuristics for LambdaRAM prefetching.



[Leigh, UIC]

NCMIR - microscopy  
(2800x4000 24 layers)

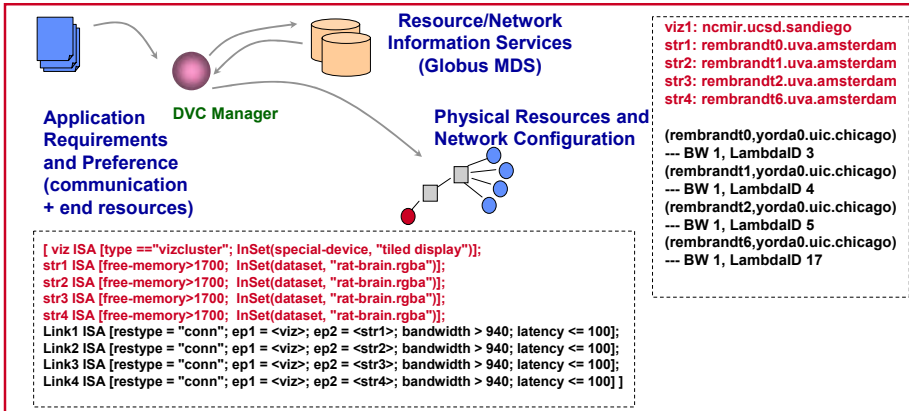


Scripps -  
Bathymetry and  
digital elevation

OptiPuter All Hands Meeting 2004 - Visualization & Data Working Group

# JuxtaView and LambdaRAM on DVC Example

## (1) Requests a Viz Cluster, Storage Servers, and High-Bandwidth Connectivity



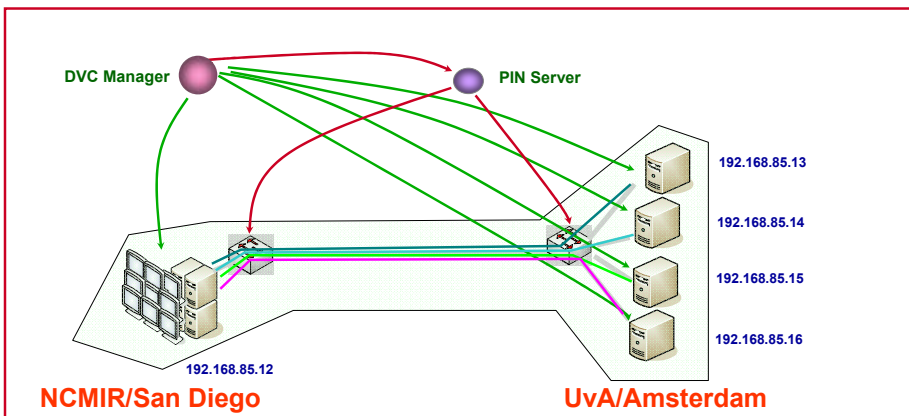
Mardi Gras -- February 4, 2005

18

# JuxtaView and LambdaRAM on DVC Example

## (2) Allocates End Resources and Communication

- Resource Binding (GRAM)
- Lambda Path Instantiation (PIN)
- DVC Name Allocation



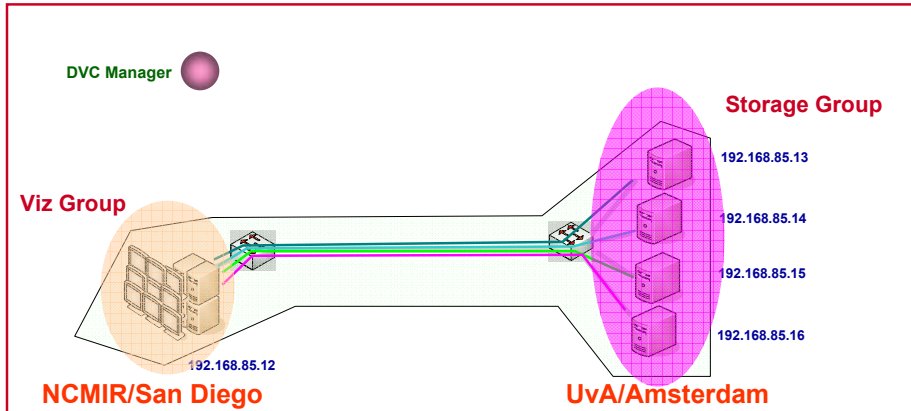
Mardi Gras -- February 4, 2005

19

# JuxtaView and LambdaRAM on DVC Example

## (3) Create Resource Groups

- Storage Group
- Viz Group



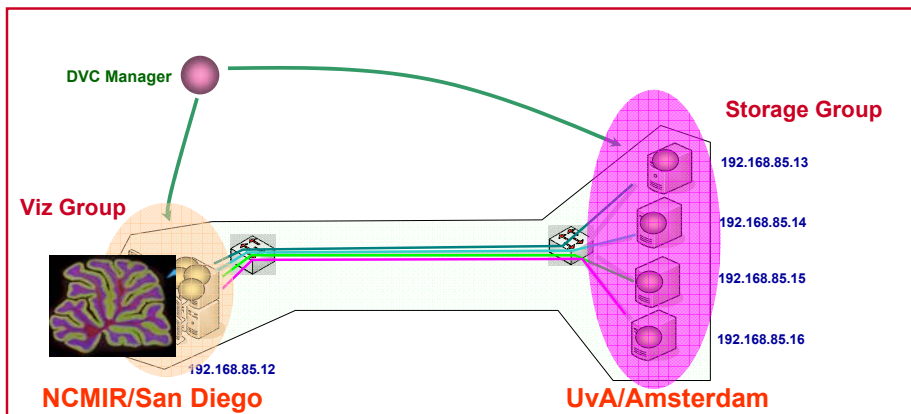
Mardi Gras -- February 4, 2005

20

# JuxtaView and LambdaRAM on DVC Example

## (4) Launch Applications

- Launch LambdaRAM Servers
- Launch JuxtaView/ LambdaRAM Clients



Mardi Gras -- February 4, 2005

21

# DVC Advantages

- **Applications**
  - **Simplifies View, Hides Complexity**
    - Automates compute/data resource binding
    - Automate dynamic  $\lambda$ -configuration; expose novel  $\lambda$ -capabilities
    - Controllable, Secure, Trusted Environment (direct access)
  - **Aggregates Resources with SAN-like model**
    - Trusted and Secure Environment
    - High Bandwidth, Deterministic (10Gbps+, no jitter)
    - Multi-party Communication
  - **Interactive, Real-time Applications**
- **System**
  - **Enables Optimized Resource Selection and Use**
    - Declarative Specification of Resource and Network Configuration
    - Optimized End Resource, Dedicated Lambda, and Switch Selection
    - Coordinated End Resource and Network Binding
- **Pragmatics**
  - Leverages VPN and Typical Grid Distributed Application structure
  - Incremental Deployability (VPLS, MPLS, Lambda's, etc.); Easy Integration

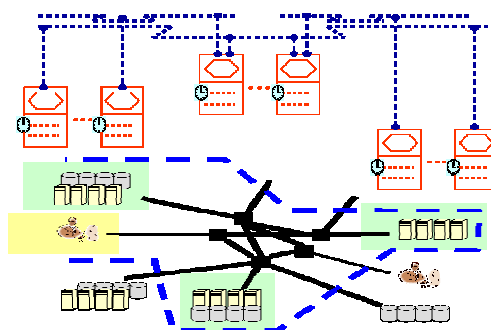
Mardi Gras -- February 4, 2005

24

# Vision -- RT Tightly Coupled Wide-Area Distributed Computing

**Real-Time Object (TMO) network**

Dynamically formed  
**Real-Time (RT) Dist. Virtual Computer (DVC)**



**RT DVC Facilitates**

- (1) **Message communications with easily determinable tight latency bounds;**
- (2) **Computing node operations enabling easy guaranteeing of timely progresses of threads toward computational milestones.**

Mardi Gras -- February 4, 2005

[Kane Kim, UC Irvine]

25

## Related Work

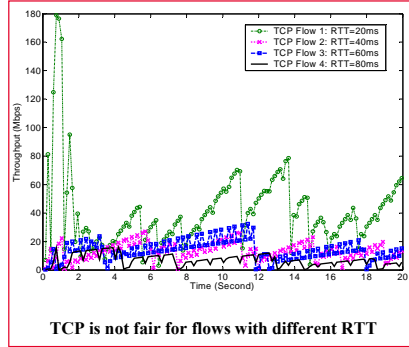
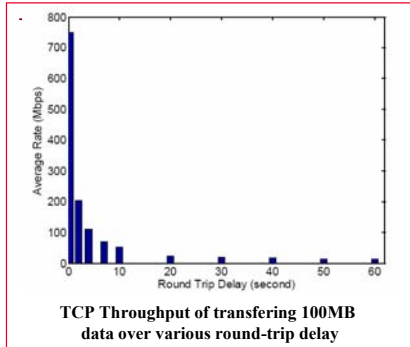
- **High Speed Optical Networks**
  - Router-based, Shared Internets
  - Flow-based Recognition
  - Transfer Based (Cheetah) [Veeraraghavan,Feng2003]
- **Distributed Application Abstractions and Tools**
  - PVM [Geist94], MPI [Various]
  - Middleware: Globus, OGSA
  - Grid Programming Tools: GridRPC [Nakada02], MPICH-G2 [Karonis03], Condor-G [Frey01], GrADS [Berman01], GridLab [Allen03]
  - Virtual Organization (VO) [Foster01]
  - Distributed Resource Context (Web Services with WSRF)
- **Distributed Virtual Computer Provides an Application-Focused Dynamic Resource Container**
  - Dynamic resource configuration and sharing policies

## OptIPuter System Software Challenges

- **What is the Model of Use for Dynamic Lambdas?**
- **How Do We Exploit the Communication Capabilities of Lambdas?**
  - High Bandwidth-Delay Product Networks
  - Endpoint Congestion (GTP)
  - Flows Faster than End Devices (CEP)
- **How Do We Support Emerging Data-intensive Applications?**

# High Performance Transport Problem

- **Growing Gap Between High Speed Links and Delivered Application Performance**
- **Transport Protocols Are a Weak Link**
  - TCP has Problems in High Bandwidth Delay Product Networks
  - “Private” Lambda-Grid Networks have new Properties
- **Efficient Point-to-Point: TCP Variants and Rate-based protocols**
- **Efficient Multipoint-to-Point**

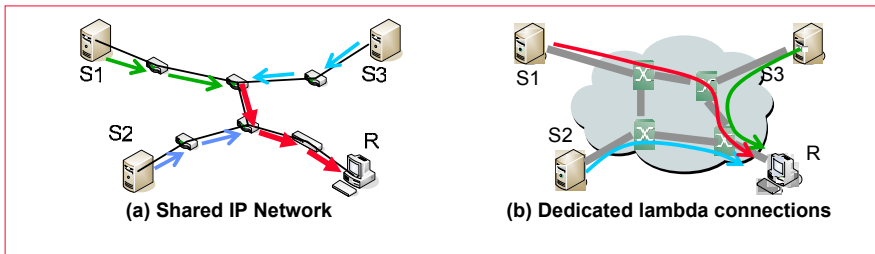


Mardi Gras -- February 4, 2005

28

# Optical Network Cores Shift Contention to Network Edge

- **Lambda-Grid: Dedicated Optical Connections Provide Plentiful Core Bandwidth**
- **Driving Applications Access Many High Data Rate Sources**
  - Multipoint-to-point communication
- => Congestion moves to the endpoints
- **Group Transport Protocol: Rate-based + Receiver Based Management**



Mardi Gras -- February 4, 2005

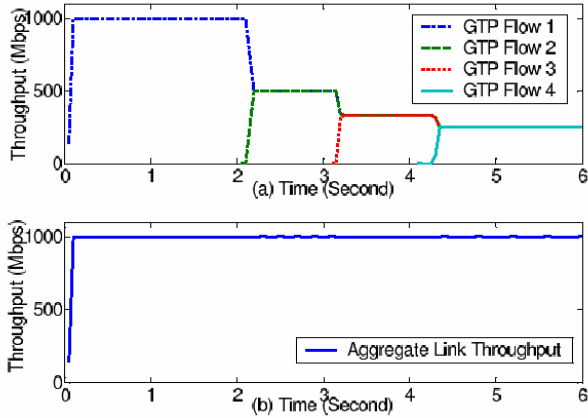
[Wu & Chien, UCSD]

29

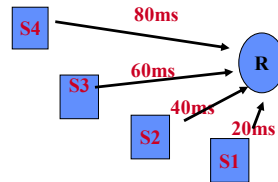
# Fairness and Convergence

- Multipoint Performance in NS2 Simulations**

- Four GTP flows with RTT 20, 40, 60 and 80ms starting at time 0, 2, 3, and 4s.



**Converging Flows:**



- GTP uses Receiver-based Management to achieve Rapid Convergence and Fair Allocation**

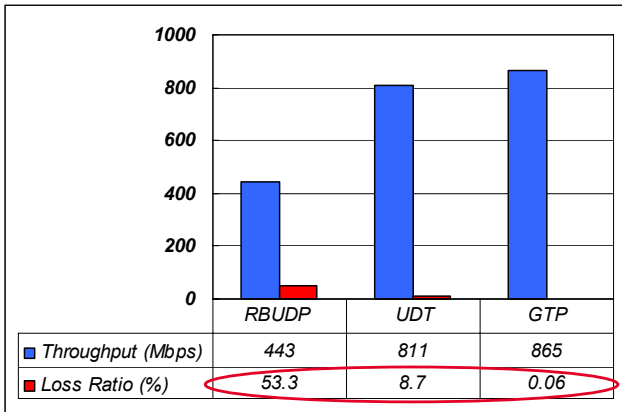
Mardi Gras -- February 4, 2005

[Wu & Chien, UCSD]

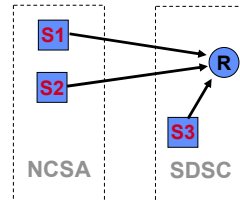
32

# Benefits of Receiver-Based Control

- SDSC -- NCSA, 10GB transfer (1Gbps link capacity), 58ms RTT
- Convergent Flows
- GTP outperforms the other Rate-based Protocols due to Receiver-oriented management



**Converging flows:**



Mardi Gras -- February 4, 2005

34

## Related Work

- **High Speed Protocol Work for Shared IP Networks**
  - HSTCP [Floyd2002]
  - XCP [Katabi2002] and Implementations [USC ISI ], ECN []
  - FAST TCP[Jin2004]
  - drsTCP[Feng2002]
- **Rate Based Protocols**
  - NETBLT, satellite channels [Clark87]
  - RBUDP on Amsterdam—Chicago OC-12 link [Leigh2002] & LambdaStream
    - For QoS enabled environment, no rate adaptation scheme
  - SABUL/UDT [Grossman2003,2004]
    - End-to-end, a combination of several control schemes.
  - Tsunami: File transfer, disk-to-disk
- **GTP Is A Novel Rate-based Protocol**
  - Employs Receiver-driven Congestion Management
  - Achieves Excellent Single And Multi-flow Performance

## Composite-EndPoint Communication

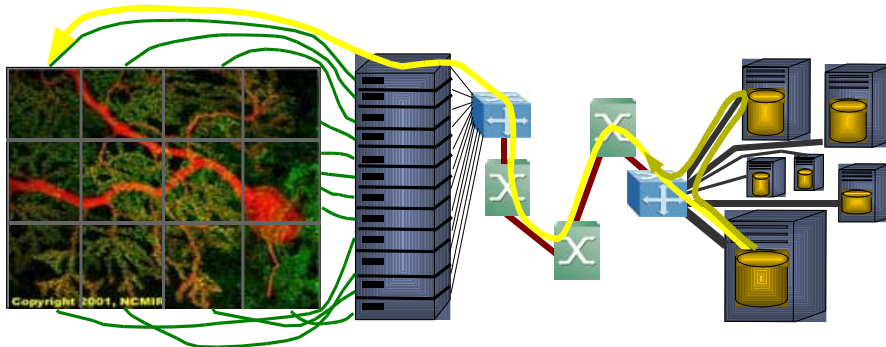


Uh-oh!

- **Network Transfers Faster than Individual Machines**
  - A 10Gbps flow? 100Gbps? A Terabit flow?
  - Use Clusters as Cost-effective, Scalable means to terminate Fast transfers
  - Support Flexible, Robust, General *N-to-M Communication*
  - Manage Heterogeneity, Multiple Transfers, Data Accessibility; Automatically

# N-to-M Example

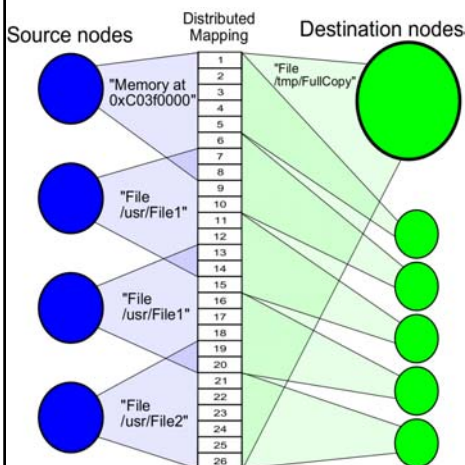
- Move Data from a Heterogeneous Storage Cluster (N)
- Exploit Heterogeneous network structure and Dedicated Lambda's
- Terminate in a Visualization Cluster (M)
- Render for a Tiled Display Wall (M)
  - Node and Pairwise Transport Properties Vary (statically, dynamically)
  - Mixed Node Memory and Storage Sources and Sinks



Mardi Gras -- February 4, 2005

37

# Composite Endpoint Protocol (CEP)

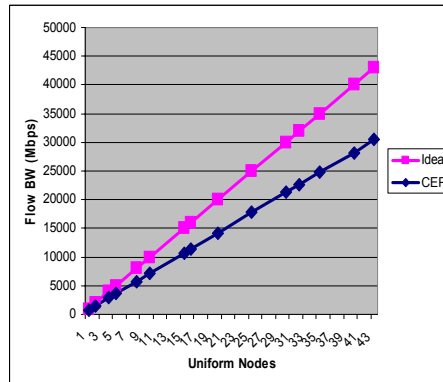
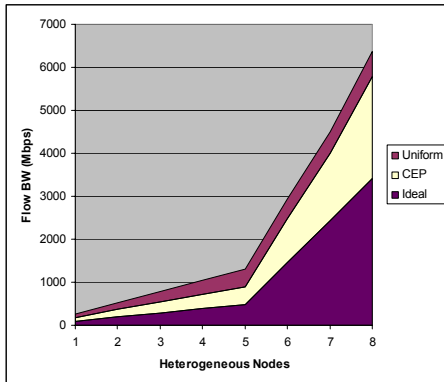


- Transfers Move Distributed Data
  - Provides hybrid memory/file namespace for any transfer request
- Choose Dynamic Subset of Nodes to Transfer Data
  - Performance Management for Heterogeneity, Dynamic Properties Integrated with Fairness
- API and Scheduling
  - API enables easy use
  - Scheduler handles performance, fairness, adaptation
- Exploit Many Transport Protocols

Mardi Gras -- February 4, 2005

38

## CEP Efficiently Composes Heterogeneous and Homogeneous Cluster Nodes



- **Seamless Composition of Performance, Widely Varying Node Performance**
- **High Composition efficiency, demonstrated 32Gbps from 1Gbps nodes!**
  - Efficiency Increasing as Implementation Improves
  - Scaling extrapolation suggests 1000-node Composite Endpoints are Feasible

Mardi Gras -- February 4, 2005

39

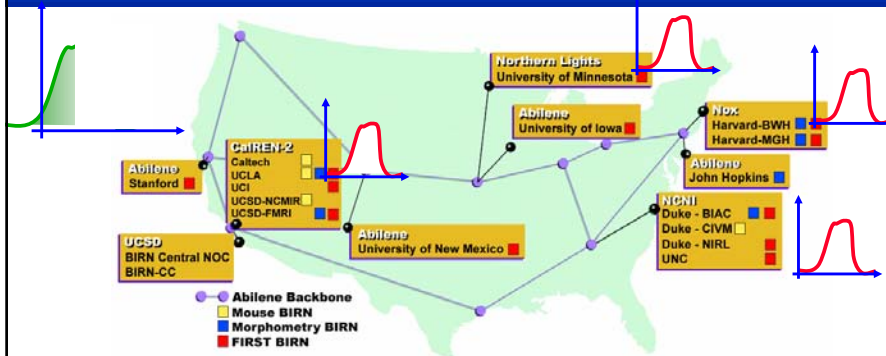
## OptIPuter System Software Challenges

- **What is the Model of Use for Dynamic Lambdas?**
- **How Do We Exploit the Communication Capabilities of Lambdas?**
- **How Do We Support Emerging Data-intensive Applications?**
  - **RobuSTore: Robust Access to Shared Disks**

Mardi Gras -- February 4, 2005

40

# RobuStore: Gigabytes per Second from Geographically Distributed Storage

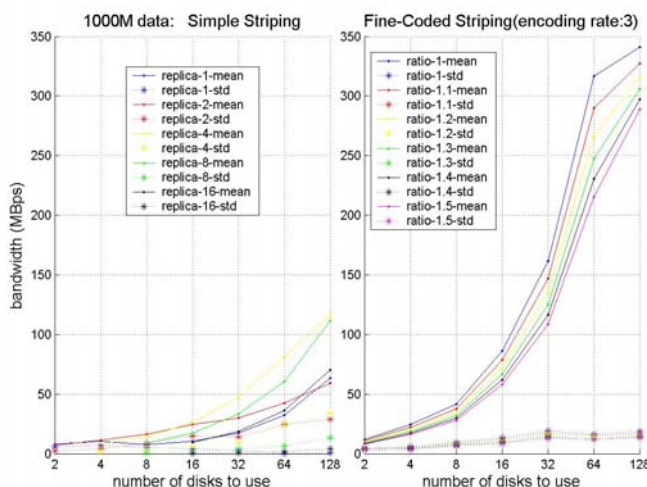


- **BIRN: Distributed Data, Intensive Analysis: 100GB – 1 PB**
  - Comparative and Collective Analysis, Visualization of Multi-Scale Data Objects
  - How to Access Data From Many Devices and Sites with High Performance?
  - How to Share the Devices and Sites with Good Performance?
- **RobuStore: Statistical Storage**
  - Systematic Introduction of Redundancy, High Efficiency LDPC Codes
  - Improve Aggregate Statistical Properties of Access => Better Performance
  - High Parallel Performance, Isolatable Performance in Shared Environments

Mardi Gras -- February 4, 2005

41

# Preliminary RobuStore Simulation Results



**Disks: Same Type, Different Layout**

**Simple Striping: 1-16x Storage Overhead**

**Erasure Code: 3x Storage Overhead**

- **Read 1GB Data: Simple Striping versus Erasure-Coded Striping**
  - RobuStore use of Erasure Codes Improvement
  - 3-5x Average Performance
  - 3x Standard Deviation

Mardi Gras -- February 4, 2005

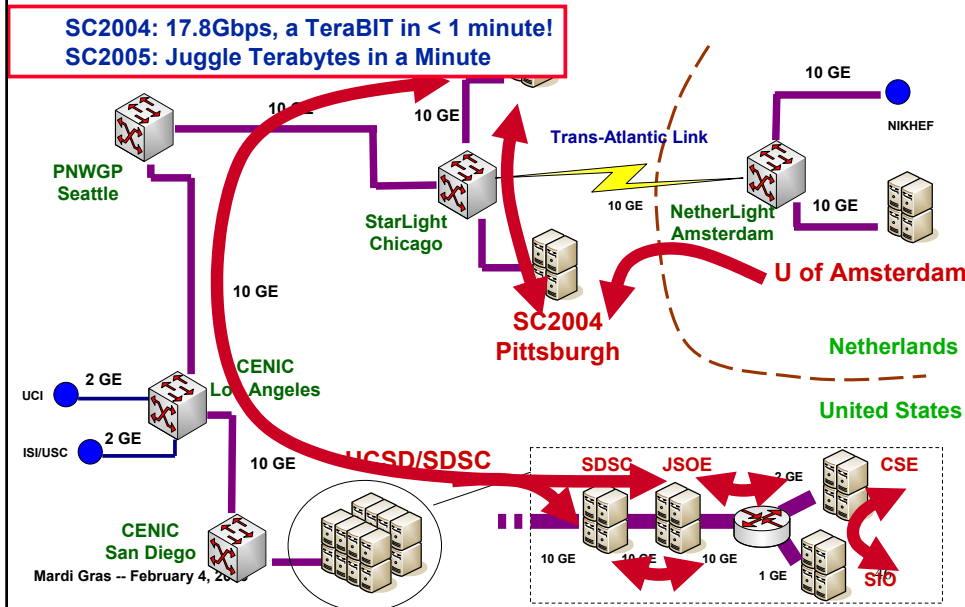
43

# Demonstrations

## Usable Terabit Networks

- Networks of many 10Gbit and 40 Gbit Links
- TeraBIT Juggling [SC2004, November 8-12, 2004]
  - Move data between OptiPuter Network Endpoints (UCSD, UIC, Pittsburgh)
  - Share efficiently; Provide Good Flow Behavior
  - Maximize Overall Transfer Speeds (all receivers saturated)
  - Configuration
    - Distributed Virtual Computer (DVC) organizes underlying Grid resources
    - Group Transport Protocol (GTP) – manages multiple converging flows
    - 10 endpoints, 40+ nodes, 1000's of miles
    - Many Converging and Crossing Flows
  - Achieved 17.8Gbps, moved a TeraBIT in less than one minute!

# 10Gig WANs: Terabit Juggling



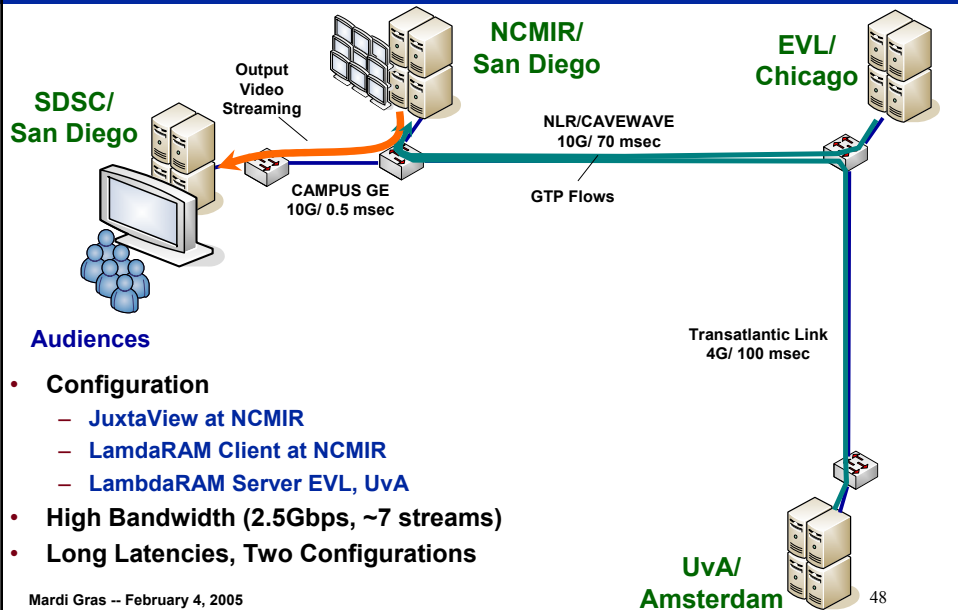
## 3-layer Integrated Demonstration

Nut Taesombut, Venkat Vishwanath, Ryan Wu, Freek Dijkstra,  
 David Lee, Aaron Chin, Lance Long  
 UCSD/CSAG, UIC, UvA, UCSD/NCMIR, etc.

January 2005, OptIPuter All Hands Meeting

1. Visualization Application (Juxtaview + LambdaRAM)
2. System SW Framework (Distributed Virtual Computer)
3. System SW Transports (GTP, UDT, etc.)

## 3-Layer Demo Configuration



## Summary and Future Work

- **Optical Networks change the balance of Distributed Systems and Grids**
  - OptIPuter is a prototype of these future capabilities
- **OptIPuter System Software delivers these capabilities to Applications**
- **Distributed Virtual Computers: Simple Collective Resource Abstraction**
  - Naming, Groups, Security, P-to-P Communication, Collective, Storage
- **Lambda's + New Transports -> Terabit Networks**
  - Group-Transport Protocol (GTP): Delivers High Speed Flows, Converging Flows, Fairness with varied RTT
  - Composite Endpoint Protocol (CEP): Flows Faster than Computers, Composition of Large #'s of Resources
- **OptIPuter: Much More to Come!**
  - Integrated Demonstrations – Real Applications and Testbeds
  - Large-scale Use of Novel Network Protocols: TeraByte Juggling
  - Large Scale Aggregate Flows: Terabit Flows
  - RobuStore: Robust Direct-Access Wide-Area Storage

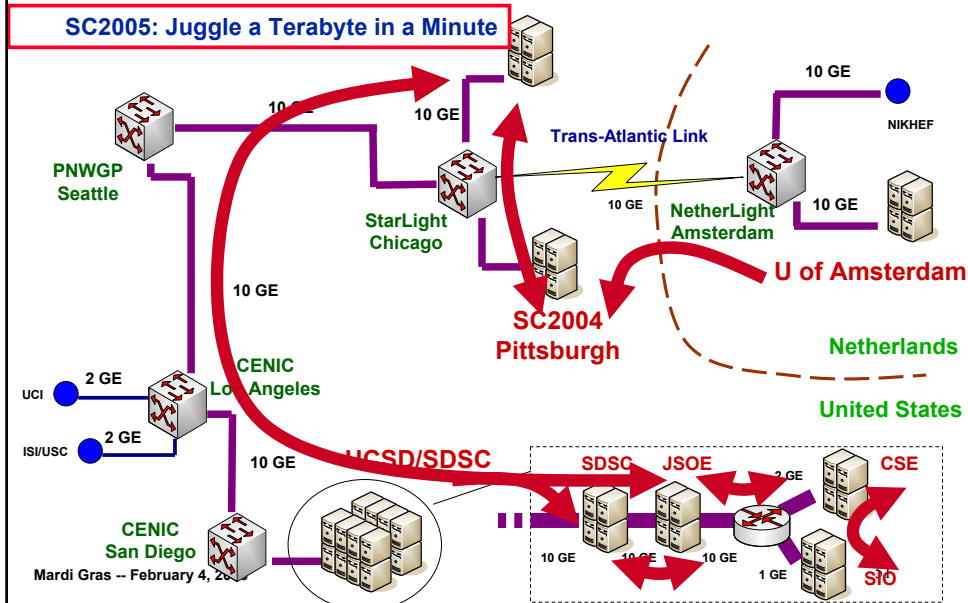
# 5-layer OptIPuter Integrated Demonstration Planned for iGrid, Sept 2005

1. Neuroscience Remote Data Access and Display (Ellisman)
2. Visualization Application (Juxtaview + Lambdaram, Leigh)
3. System SW Framework (Distributed Virtual Computer)
4. System SW Transports (GTP, UDT, etc.)
5. OptIPuter Distr. Optical Backplane Software - PIN/ODIN (Mambretti/Yu)

Mardi Gras -- February 4, 2005

50

# Terabyte Juggling: SC2005?



# UCSD/CSAG OptIPuter Team

## Faculty

- **Andrew A. Chien**

## Graduate Students

- **Network Protocols**
  - Xinran (Ryan) Wu, Eric Weigle
- **Storage**
  - Huaxia Xia, Justin Burke, Frank Uyeda
- **DVC**
  - Nut Taesombut
- **Web site:** <http://www-csag.ucsd.edu/>

# For More Information

- **Distributed Virtual Computers, System Software Model**
  - L. Smarr, A. Chien, T. DeFanti, J. Leigh, P. Papadopoulos, [The OptIPuter](#), *Communications of the Association for Computing Machinery (CACM)*, 47(11), November 2003.
  - N. Taesombut and A. Chien, [Distributed Virtual Computer \(DVC\): Simplifying the Development of Grid Applications](#), *Grids and Advanced Networks (GAN) at CCGrid 2004*, April 2004
  - Andrew A. Chien, Xinran (Ryan) Wu, Nut Taesombut, Eric Weigle, Huaxia Xia, and Justin Burke, [OptIPuter System Software Framework](#), *UCSD Technical Report CS2004-0786*.
  - Kane Kim, [Wide-Area Real-Time Distributed Computing in a Tightly Managed Optical Grid - An Optiputer Vision](#), Paper and Keynote speech at *Advanced Information Networking and Applications 2004*, Fukuoka, March, 2004.
- **OptIPuter Transport Protocols**
  - E. Weigle and A. Chien, [The Composite Endpoint Protocol \(CEP\): Scalable Endpoints for Terabit Flows](#), *IEEE Symposium on Cluster Computing and the Grid*, April 2005, Cardiff, United Kingdom.
  - X. Wu and A. Chien, [GTP: Group Transport Protocol for Lambda Grids](#), *IEEE Symposium on Cluster Computing and the Grid (CCGrid)*, April 2004.
  - X. Wu and A. Chien, [Evaluation of Rate-based Transport Protocols for Lambda Grids](#), *IEEE Conference on High-Performance Distributed Computing (HPDC-13)*, June 2004
  - Y. Gu and R. Grossman, [Optimizing UDP-based Protocol Implementations](#), *Third Workshop on Protocols for Long Distance Networks (PFLDNet)*, February 2005.
  - A. Falk, T. Faber, J. Bannister, A. Chien, R. Grossman, J. Leigh, [Transport protocols for high performance](#), *Communications of the ACM*, Volume 46, Number 11, November 2003, pp. 42-49.

**Questions?**