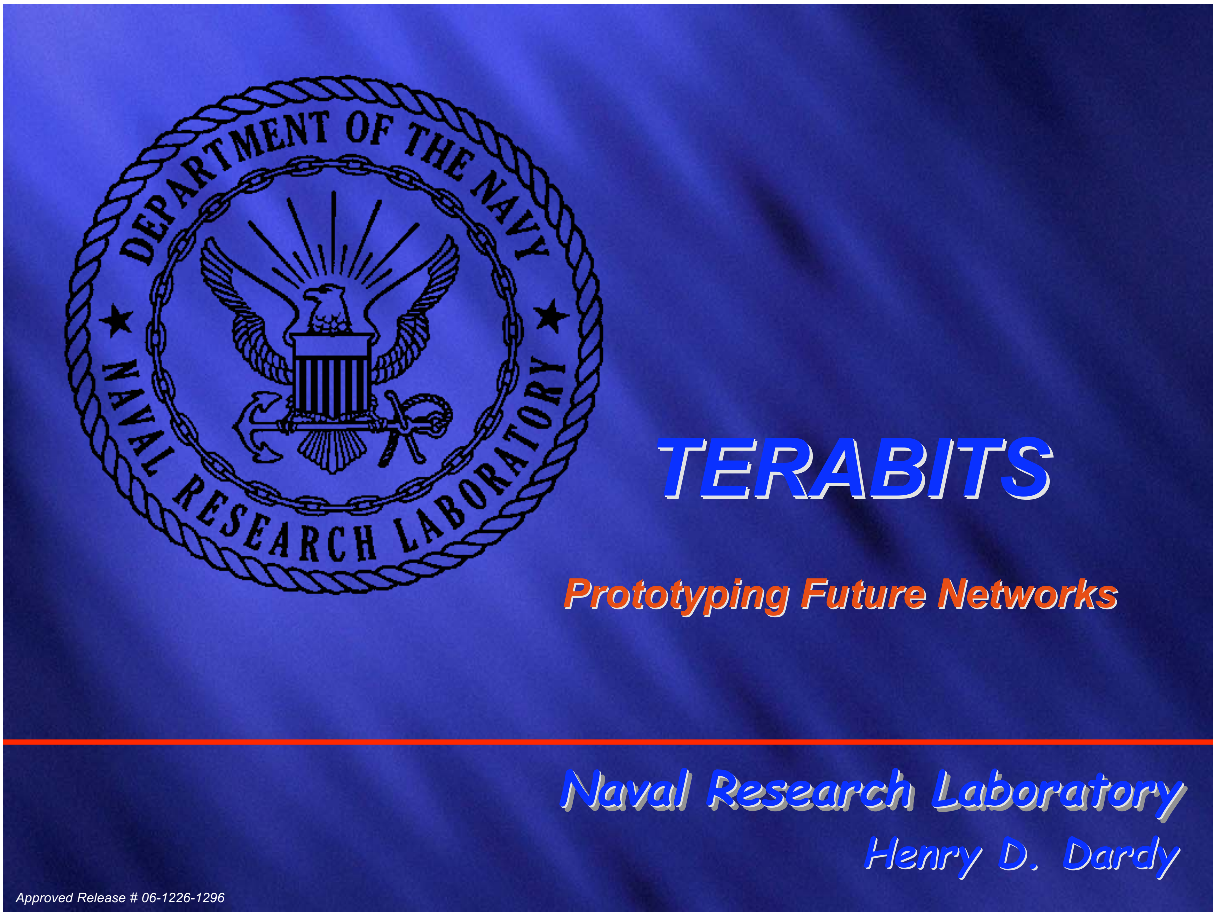


TERABITS

Prototyping Future Networks

Naval Research Laboratory

Henry D. Dardy

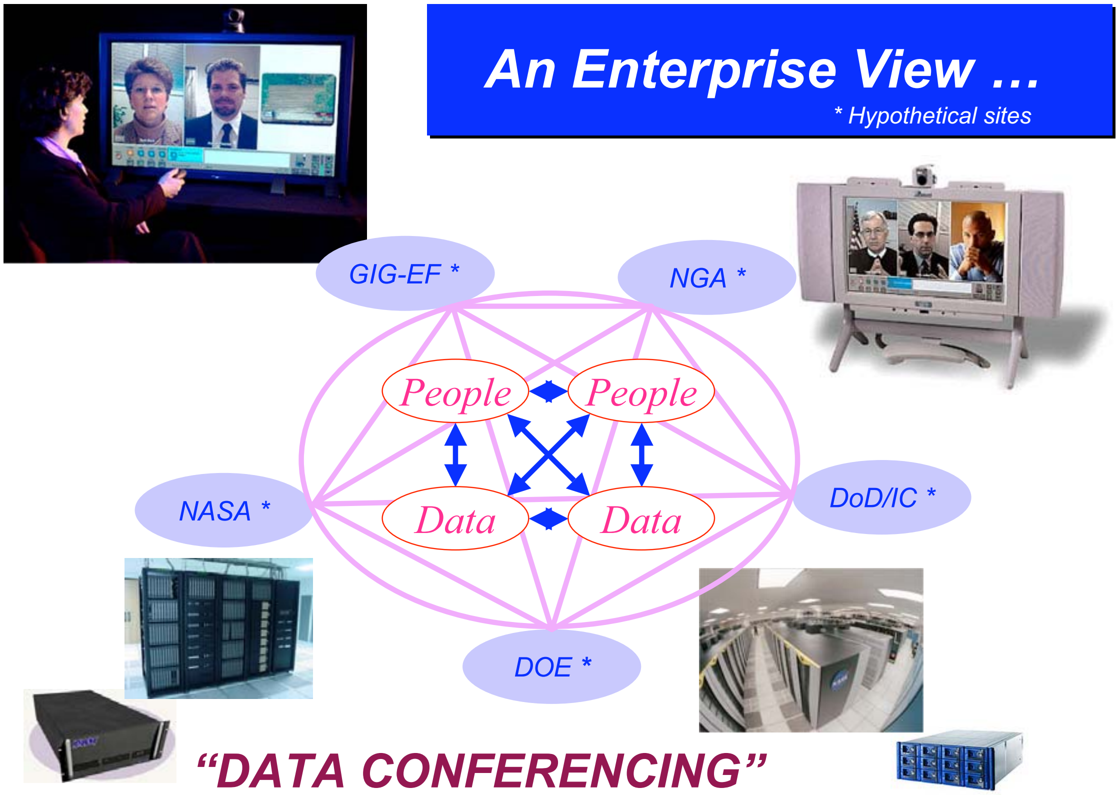
Approved Release # 06-1226-1296

A Terabit Challenge . . .

Build a Global "Large Data" Network Infrastructure to Rapidly Access and Produce Knowledge from the Best Information available from Federated, Distributed sensors and digital media assets

- Integrate federated, distributed computational grids, realtime sensors, and digital historical information
- Scale to support exponentially increasing data archives
- Privacy, authenticity and security demands: InfoAssured
- Affordable … highly available … E2E QoS flows
- Legacy and rapidly evolving technology integration
- Perf, NetOps, Information Assurance tools/sensors
- Reachback, Traceback realtime capabilities

"Expose interfaces early and often!"



An Enterprise View …

* Hypothetical sites

“DATA CONFERENCING”

… multiple sites, people (O2M), P2P seamlessly interacting!

big *fast* “*terabytes/hour*” data problem …

… efficiently interface high performance optical networks directly to

- *Supercomputers*
- *Grid Clusters*
- *Visualization, SuperHDTV*
- *HR Motion Imagery*
- *TSAT Tactical Comms*
- *GIS Imagery/Weather/Oceans*
- *2D/3D workstations*
- *Online Digital Asset Archives*
- *Hyperspectral …40K x 40K*
- *Virtualized Ground Station*

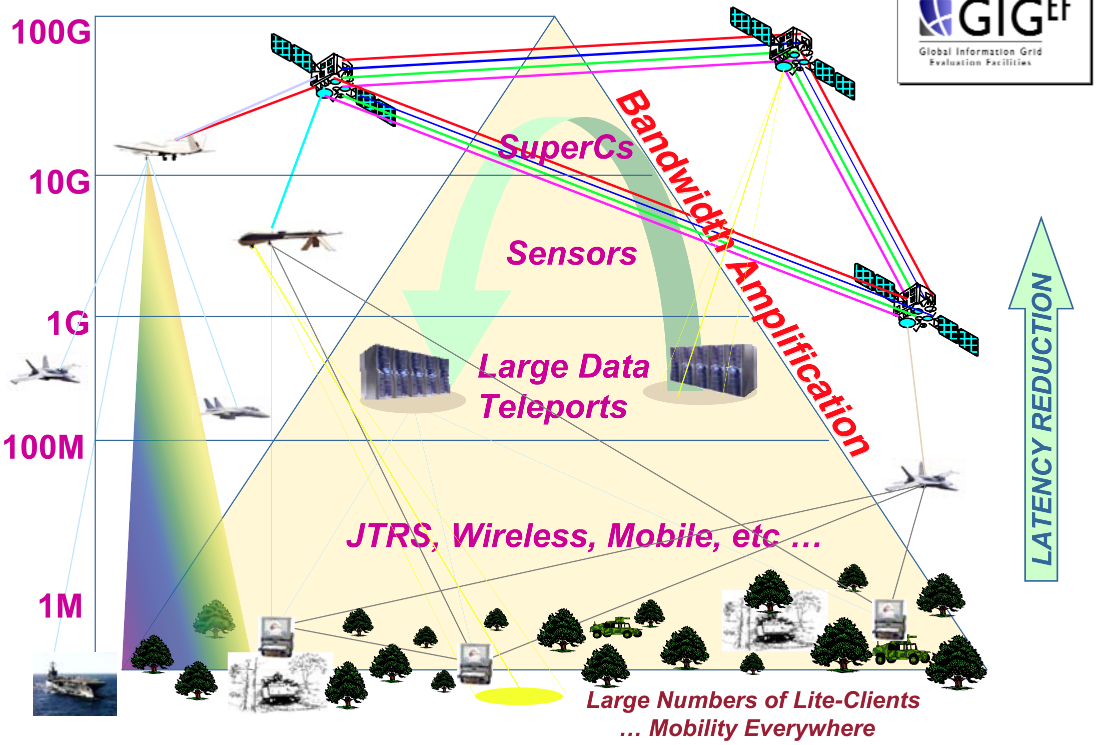
- Interfaces need to scale as *Optical LAN* networks scale
- Interface programming model and semantics familiar and friendly
- Minimum of equipment required for each *lambda* connection
- WAN transport protocol semantics simply abstracted from applications
- Sustained performance across the WAN approaches *full wire speed*
- *Multicast, QoS, MLPP, IA, Encryption* supported by protocol E2E

-Routinely exchanging multi-TByte streamed data sets long haul during daily workflows from sensors …

-Multi-PetaByte online distributed, federated archives

A Net-centric Architecture . . .

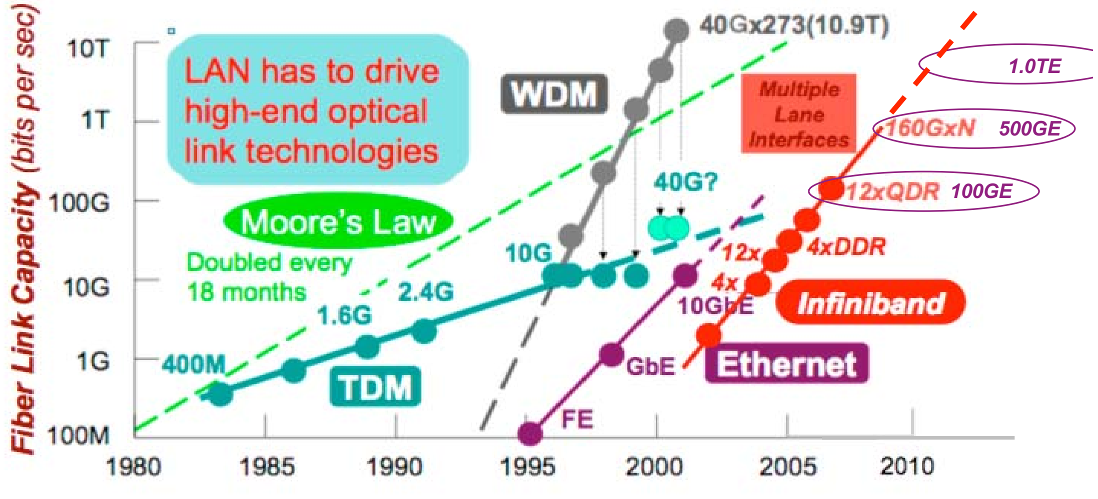
GIGEF Global Information Grid Evaluation Facilities



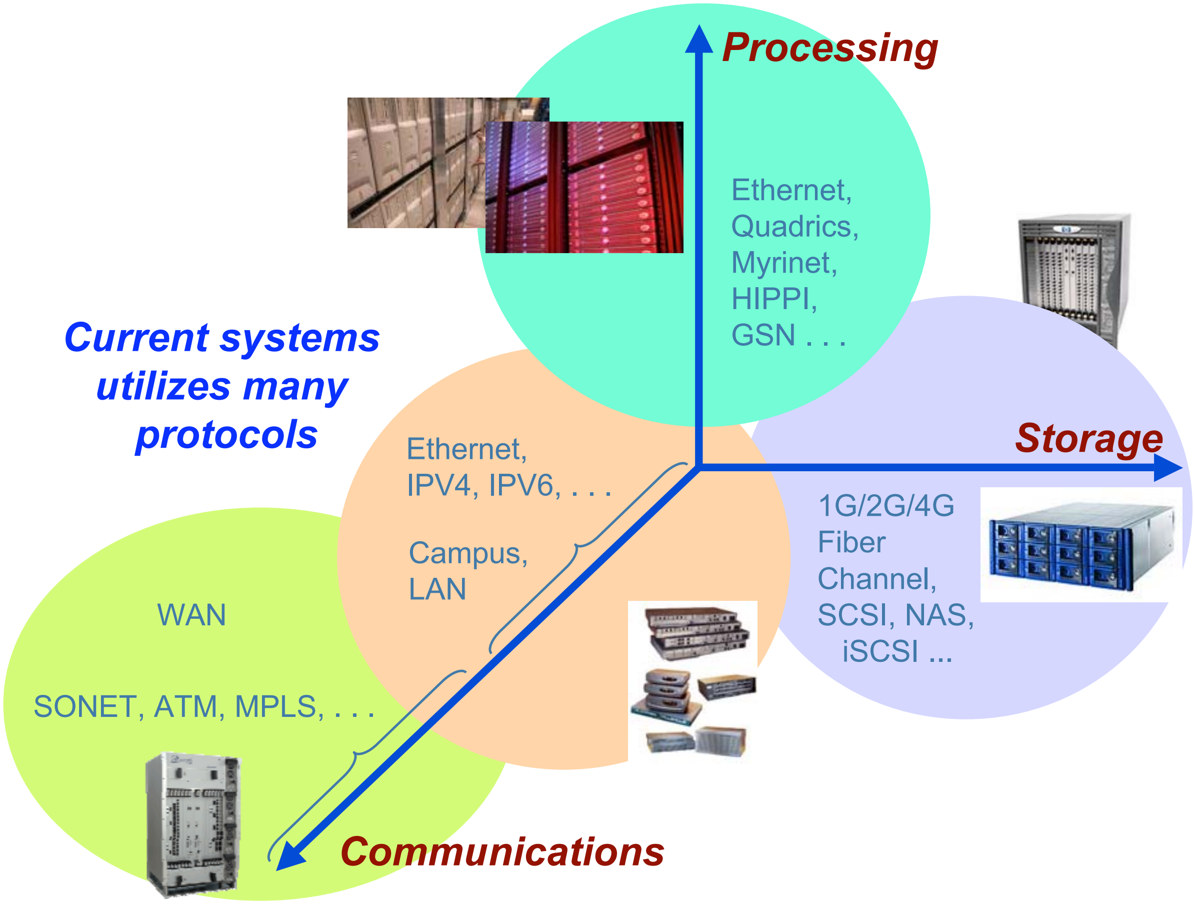
Large Numbers of Lite-Clients
… Mobility Everywhere

"… a single packet triggers High Bandwidth Flows …"

Optical Link Performance, per Laser



Ref: O. Ishida, NTT, "Toward Terabit LAN/WAN" Panel, iGrid 2005
H. Dardy, NRL, Infiniband Multiple Lane Interfaces, 100/500GE …



Current systems utilizes many protocols

Processing

Ethernet, Quadrics, Myrinet, HIPPI, GSN . . .

Storage

1G/2G/4G Fiber Channel, SCSI, NAS, iSCSI ...

Ethernet, IPV4, IPV6, . . .

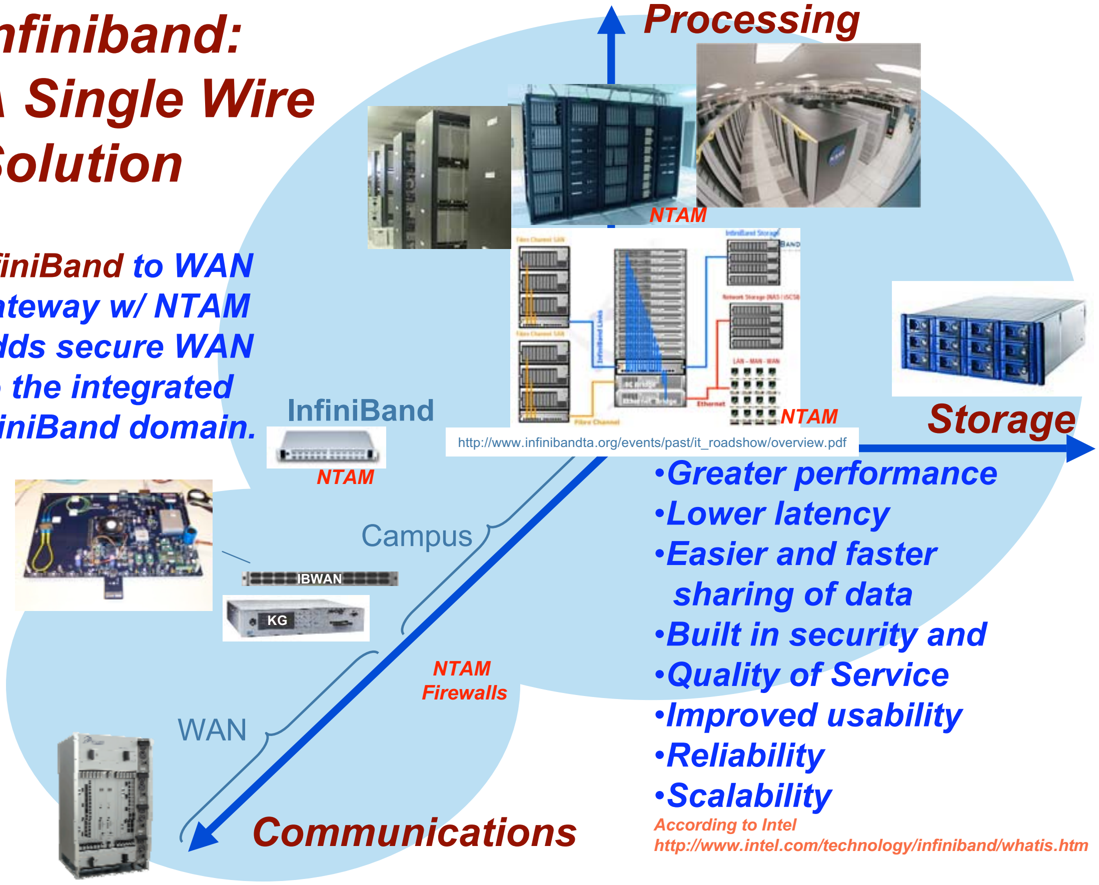
Campus, LAN

WAN

SONET, ATM, MPLS, . . .

Communications

Infiniband: A Single Wire Solution



InfiniBand to WAN Gateway w/ NTAM adds secure WAN to the integrated InfiniBand domain.

Processing

NTAM

NTAM

http://www.infinibandta.org/events/past/it_roadshow/overview.pdf

Storage

InfiniBand

NTAM

IBWAN

KG

Campus

NTAM Firewalls

WAN

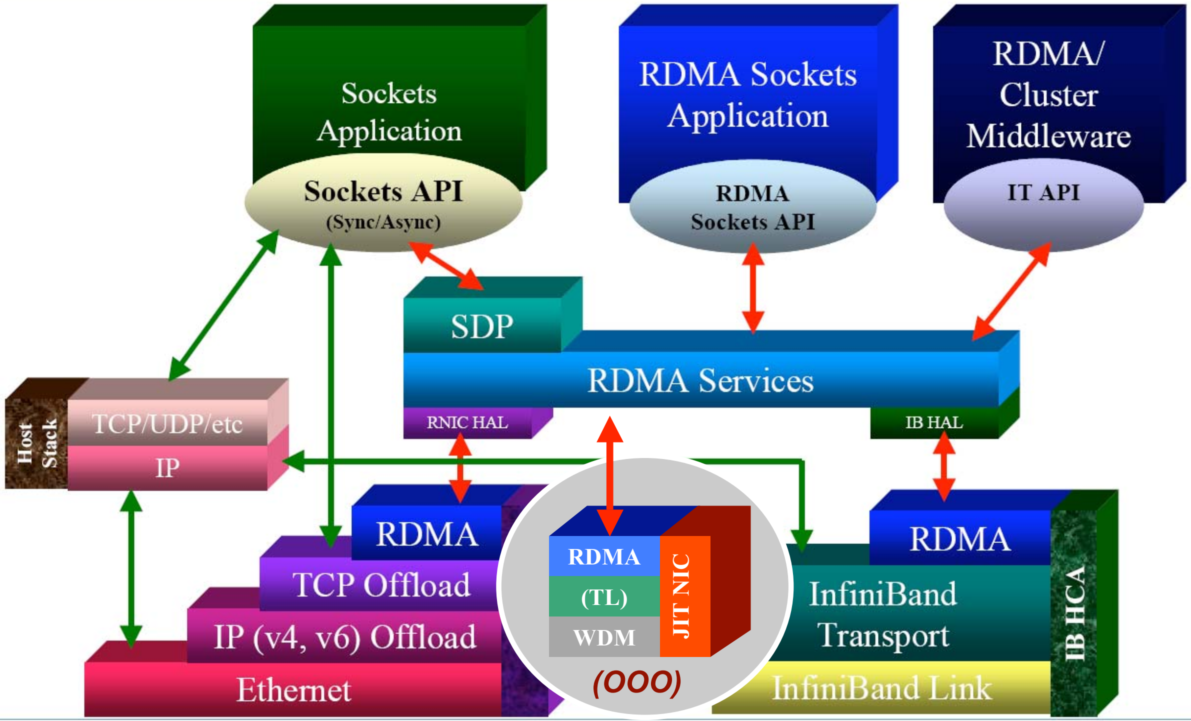
Communications

- *Greater performance*
- *Lower latency*
- *Easier and faster sharing of data*
- *Built in security and*
- *Quality of Service*
- *Improved usability*
- *Reliability*
- *Scalability*

According to Intel
http://www.intel.com/technology/infiniband/whatis.htm

RDMA Infrastructure: *Solution Components*

http://www.mellanox.com/shared/hp_ci_oracle_world.pdf



Sockets Application

Sockets API (Sync/Async)

RDMA Sockets Application

RDMA Sockets API

RDMA/ Cluster Middleware

IT API

SDP

RDMA Services

RNIC HAL

IB HAL

Host Stack

TCP/UDP/etc

IP

RDMA

TCP Offload

IP (v4, v6) Offload

Ethernet

RDMA

(TL)

WDM

JIT NIC

(OOO)

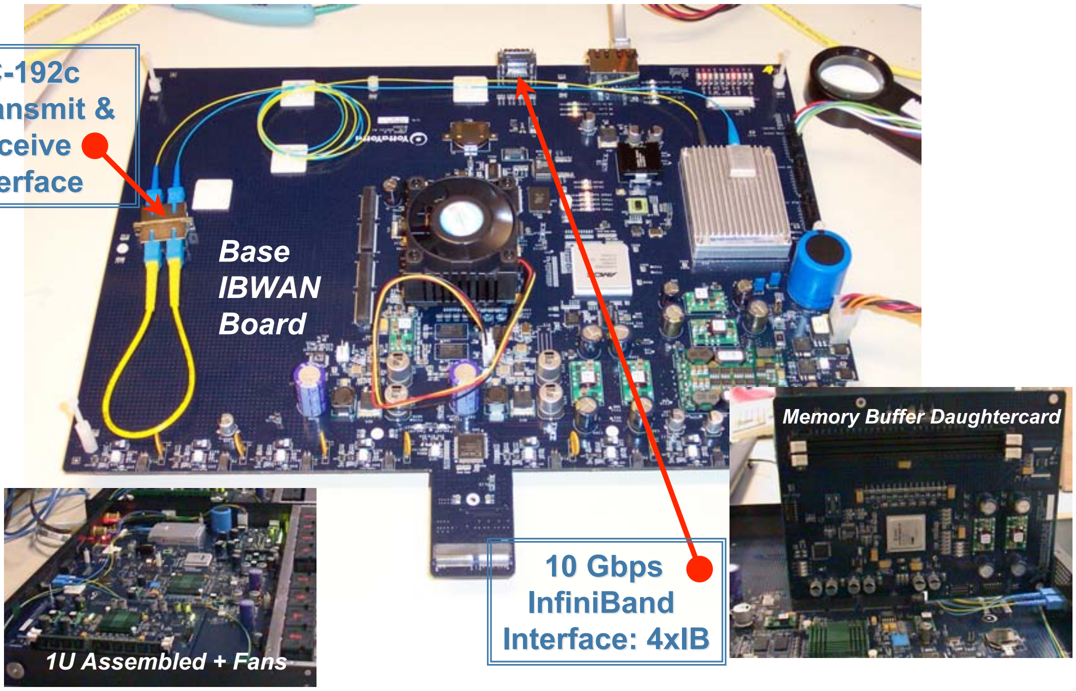
RDMA

InfiniBand Transport

InfiniBand Link

IB HCA

IBWAN: Functional Prototype …



OC-192c Transmit & Receive Interface

Base IBWAN Board

Memory Buffer Daughtercard

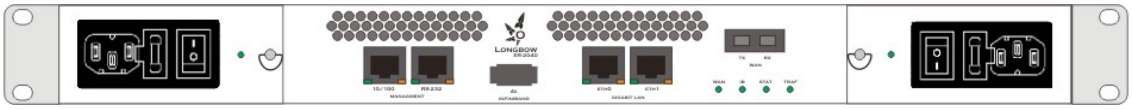
1U Assembled + Fans

10 Gbps InfiniBand Interface: 4xIB

Range Extended InfiniBand . . . Next Steps

Performs InfiniBand encapsulation over 10GE, POS and ATM WANs at 4x InfiniBand (10 Gbps, 8b/10b speeds) … *useable w/Type I Encryption*

- *Looks like a 2-port InfiniBand switch or router to the IB fabric*
- *Designed for 100,000 km+ distances for fiber or satcom links*
- *NRL collaborated with Obsidian Research Corp to develop IBWAN prototypes … flow based, "gargoyle" NTAM sensing, etc.*
- *Coupled with cache-coherent hardware support from YottaYotta, large data streaming is possible in realtime across global distances*
- *Productized versions of the 10Gbits/s 4xIB prototype ready*
- *Applications software being developed to facilitate deployment of wide area switched wavelength IB data streaming technology*
- *A second source digital hub is available from Bay Microsystems Inc*



Achieves 950+ MBytes/s sustained performance in a single logical flow ~ 4% CPU load (Opteron 242s using RDMA transport with cache-coherency) … IPv6 Packet Over SONET (for HAIPE when available) & ATM (KG-75a Encryption) modes.

Working toward Terabit Internetworking . . .

4x IB WAN . . . *CY2006*

Point-to-point:

- *ATM/SONET (OC-192c)*
- *IPv6 POS (OC-192c)*

Targeted: 3-way multicast

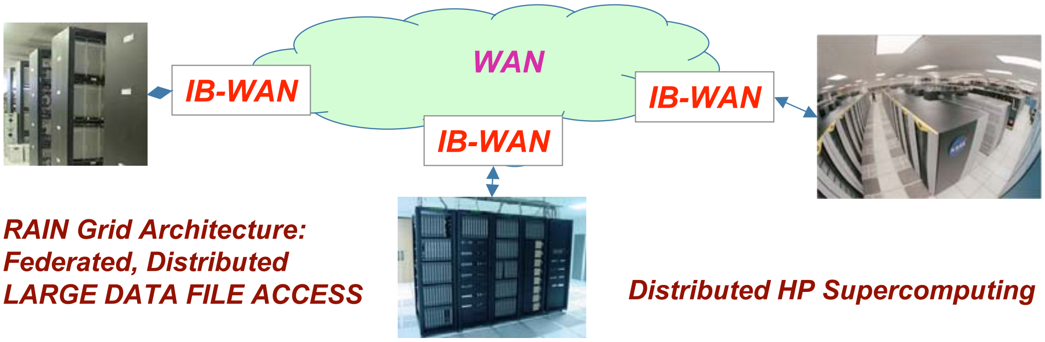
- *ATM with QOS (OC-192c or OC-48c)*
- *IPv6 POS (OC-192c or OC-48c or 10 GigE)*
- *GMPLS (preset)/ JIT (OBS research)*
- *SMPTE 292m (4:2:2 & 4:4:4) 720p/1080p*

12x DDR IB WAN

- *4Q 2006/1Q 2007*
- *GFP*
- *ATM/SONET (OC-768c)*
- *IPv6 POS (OC-768c)*
- *GMPLS (via SIP or UCLP)*
- *JIT (dynamic)*

12x QDR

- *~2008 12xQDR=100GE*



RAIN Grid Architecture: Federated, Distributed LARGE DATA FILE ACCESS

Distributed HP Supercomputing

Session Initiation Protocol . . . ***SIP***

An IETF application layer control protocol

- ***Used for establishing, manipulating, & tearing down sessions***
- ***Adopted as the VoIP and IM signaling protocol … voice, video, data, imagery … works for wavelengths with G/MPLS***
- ***Sessions viewed as a two-way call or a collaborative multi-media conference … multicast***

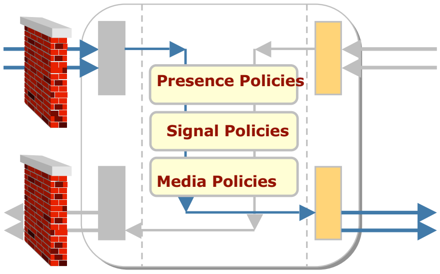
Quality of Service establishment and path selection **… policy driven**

A request-response protocol that closely resembles HTTP & SMTP

- ***Telephony (VoIP) becomes another IP web application***
- ***Alongside presence-based collaboration and real-time video***
- ***Referred to as “converged communications”***

> *“SIP is probably the third great protocol of the Internet, after TCP/IP and HTTP!”*
> ***1. Internet Communications Using SIP – Sinnreich, Johnston, John Wiley & Sons, 2003*** *… Vint Cerf*

SIP*: An Application Layer IP Control Plane*



- ***Realtime Presence Control***
- ***Multiprotocol capable***
- ***Voice, Video, Data and Imagery flows***
- ***Call Signal Control***
- ***Media Control***
- ***Signaling and Media Encryption***
- ***Protocol Validation & Intrusion Protection***
- ***Authentication & Authorization***
- ***Denial of Service Protection***

Provides a secure fabric for real-time collaboration . . .voice, video, data, imagery

- ***Single view of security & control across enterprise***
- ***Enforces corporate, group and user policies: presence, signaling & media***
- ***Federation across domains***
- ***Hardened appliance***
- ***Carrier or Enterprise scalability, availability & security***
- ***Utilization of existing infrastructure***
- ***Agnostic to transport***

What is a **GARGOYLE** sensor ?

Comprehensive Passive, Transparent Real-Time Flow Monitoring

- User Plane and Control Plane Complete Information Assured Transaction Monitoring
- Reporting on System/Network QoS status with every use
  - Capacity, Reachability, Responsiveness, Loss, Jitter
  - ICMP, ECN, Source Quench, DS Byte, TTL

Multiple Flow Strategies

- Layer 2, MPLS, VLAN, IPv4, IPv6, Layer 4 (TCP, IGMP, RTP), 4x/12x IB, …

Small Footprint

- 200K binary, sensors to supercomputers

Performance

- OC-192c, 10GB Ethernet, OC-48c, OC-12c, 100/10 MB Ethernet, SLIP
- *Ongoing research to scaling to OC-768c (40G), 100G, 160G, etc.*
- POS, ATM, Ethernet, FDDI, SLIP, PPP. Infiniband 4x/12x S/D/QDR
- > 1.2 Mpkts/sec Dual 2GHz G5 MacOS X.
- > 800Kpkts/sec Dual 2GHz Xeon Linux RH Enterprise

Supporting Multiple OS's

- Linux, Unix, Solaris, IRIX, MacOS X, Windows XP

Comprehensive Data Network Accountability

NTAM *… Provides an ability to account for all/any network use at a level of abstraction that is useful, supports all protocols, unencrypted or encrypted, at all layers and for all levels of encapsulation !*

Network Service Functional Assurance … ***End User Info***

- *Was the network service available?*
- *Was the service request appropriate?*
- *Did the traffic come and go appropriately?*
- *Did it get the treatment it was suppose to receive?*
- *Did the service initiate and terminate in a normal manner?*

Network Control Assurance … ***NetOps Info***

- *Is network control plane operational?*
- *Was the last network shift initiated by the control plane?*
- *Has the routing service converged?*

Information Assurance … ***Security and Economic Info***

- *Converged solution: network, performance, security, billing*

Network Scaling Agenda . . .

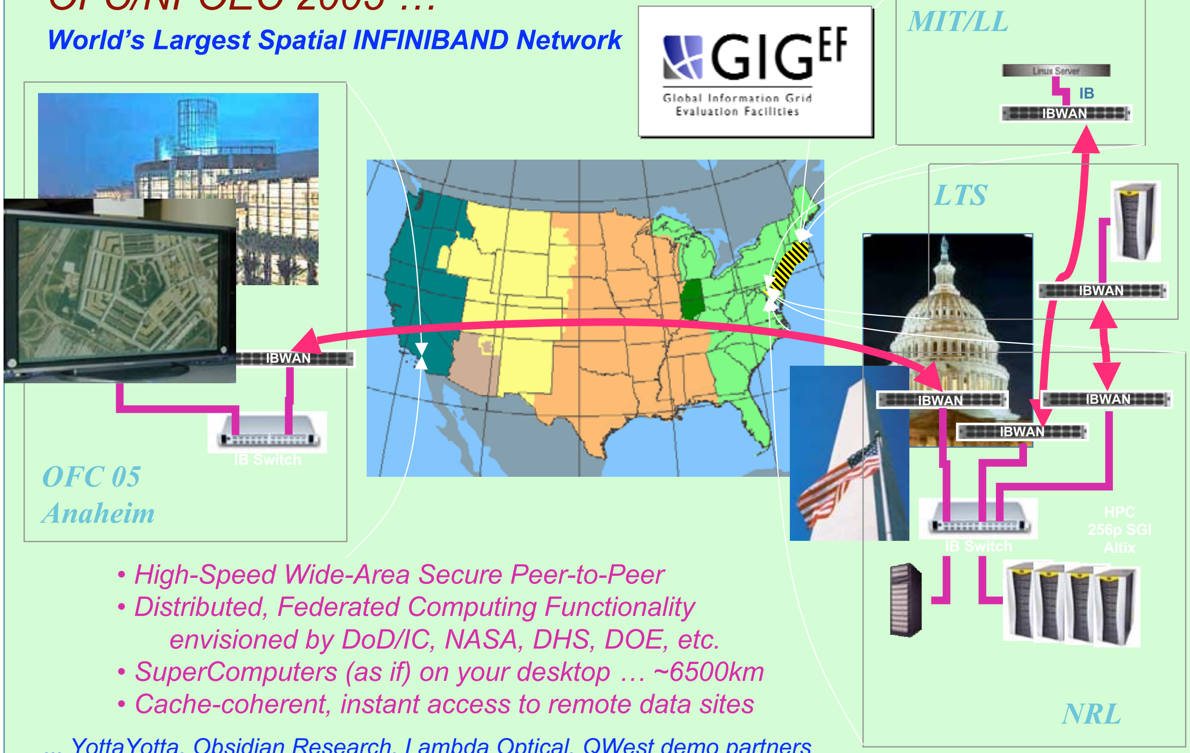
TODAY

	2005	0-2 YEARS	3-5 YEARS	5-15 YEARS
OPTICAL STREAMS	1-10 *Gbps*	10-40 *Gbps*	120-640 *Gbps*	1-10 *Tbps*
OPTICAL CNTL Plane	STATIC *Provisioned*	DYNAMIC *(GMPLS)*	BURST/JIT *Just-in-time*	
Control Plane	STATIC *Tunnel*	DYNAMIC *SIP*	SIP *QoS/QoP*	
LAN/WAN Technology	*IPV4:* 1GE, OC12c, 4xSDR Infiniband	*IPV6:* 4x/12x SDR/DDR Infbnd(cc), 10GE	*IPV6:* 12xQDR Infbnd(cc), 100GE, 64-128x IB	*All Optical System Interconnect*
SECURITY Devices	*1.0G IPV4 FW,K5,3DES, CBs, KGs, NTAM*	*10G KGs, HAIPEs, CAC, FEON, PKI, NTAM*	*40G HAIPE, Scalable GFP Encrypter*	*640G HAIPE, GFP Encptr*

SPECIAL TOPICS	*Quantum Key Distribution (QKD), Dynamic PMD Comp, Peering/Multicast, Parallel Optics, OOO(2R) Optical Regeneration, . . .*

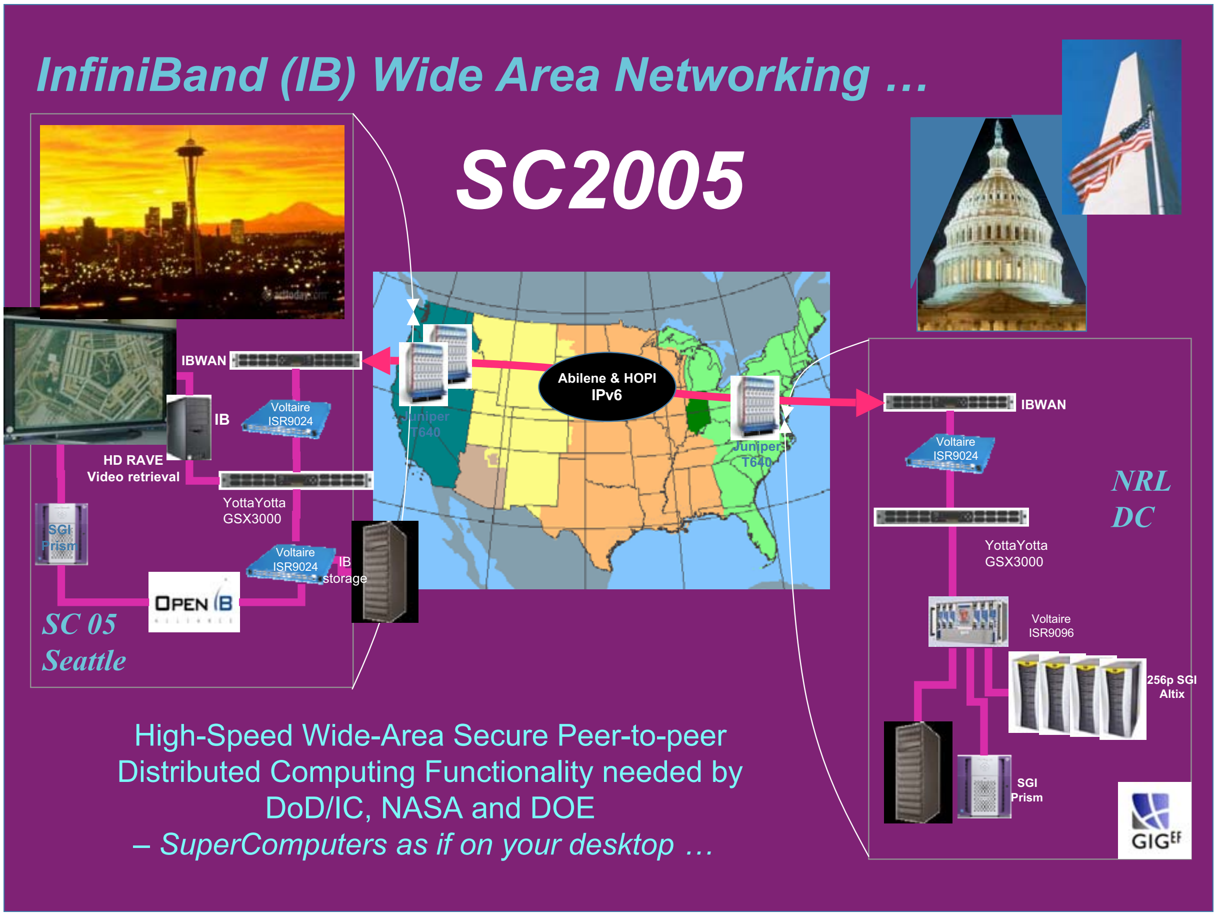
InfiniBand *Wide Area Networking*
OFC/NFOEC 2005 …

World's Largest Spatial INFINIBAND Network



- *High-Speed Wide-Area Secure Peer-to-Peer*
- *Distributed, Federated Computing Functionality envisioned by DoD/IC, NASA, DHS, DOE, etc.*
- *SuperComputers (as if) on your desktop … ~6500km*
- *Cache-coherent, instant access to remote data sites*

... YottaYotta, Obsidian Research, Lambda Optical, QWest demo partners

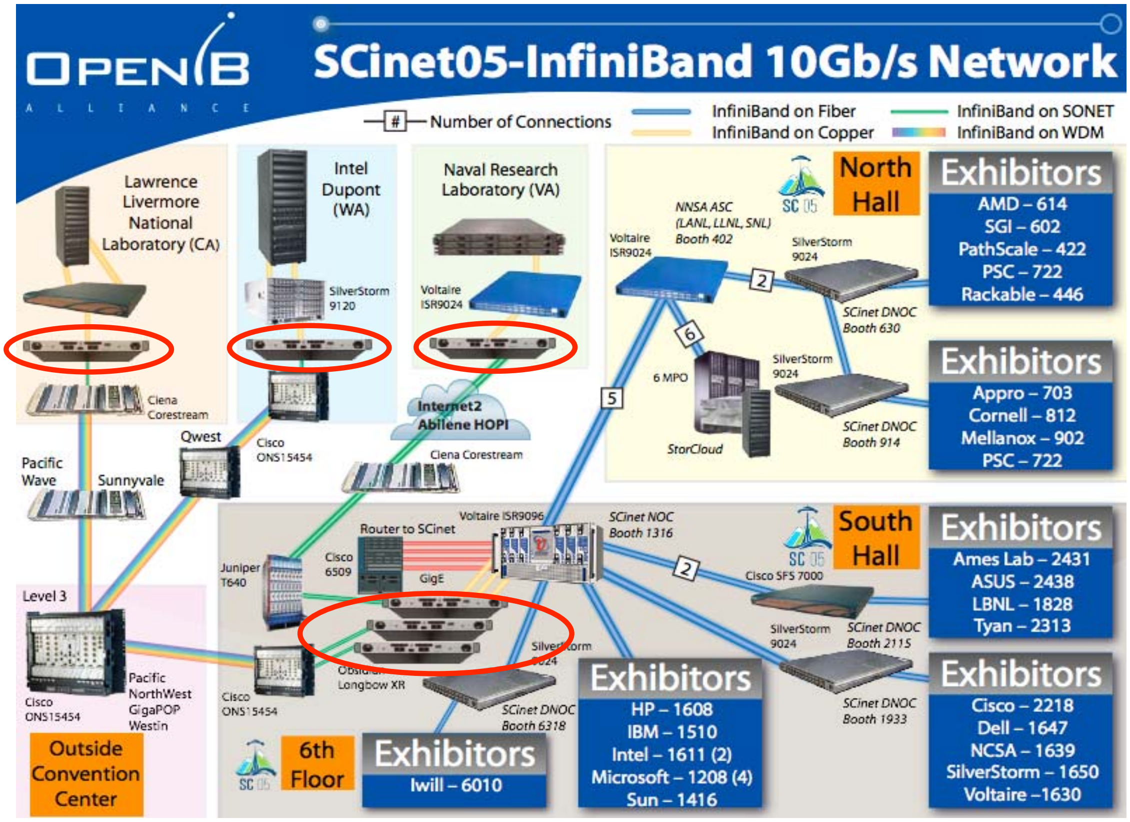


InfiniBand (IB) Wide Area Networking …

SC2005

High-Speed Wide-Area Secure Peer-to-peer Distributed Computing Functionality needed by DoD/IC, NASA and DOE

– *SuperComputers as if on your desktop …*



SCinet05-InfiniBand 10Gb/s Network

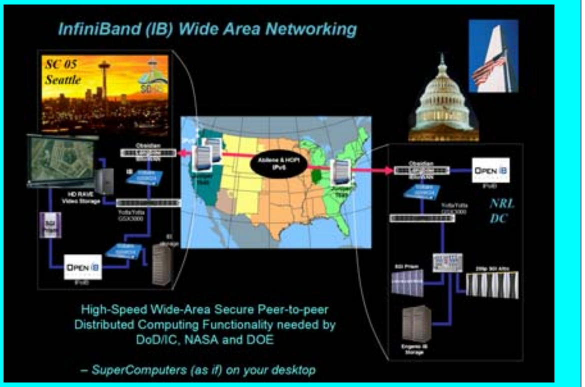
SCALING THE GLOBAL INFORMATION GRID

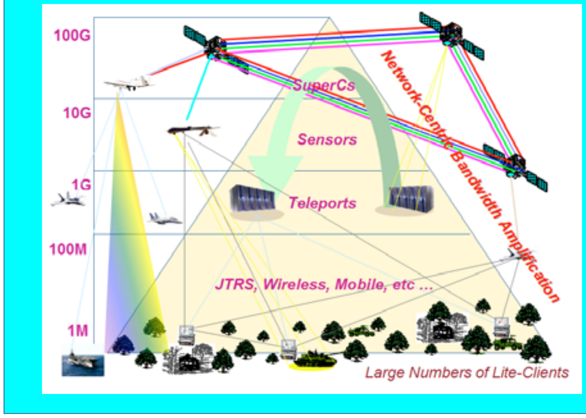
Naval Research Laboratory

Requirements to access and process large amounts of data to exploit information for knowledge have pushed the envelop of conventional architectures. The challenge for High Performance Computing and Communications is to address large data problems in a much more coordinated and rapid manner. This challenge is driven by the exponential growth in data that is driving high-end optical link technology.

Meeting this challenge requires new scalable architectural approaches. Precisely because processing needs to be coupled to distributed, federated global data and the data itself is growing at a rate significantly faster than Moore's Law, a net-centric approach must be employed that meets the conflicting needs of data locality and global consistency. This leads to defining a wholly new edge architecture that can scale to meet the challenges facing the networks in the years ahead.

The emerging ability to flexibly direct connect and securely peer sustained high stream, low-latency flows in an optimum wide area, distributed infrastructure is one of the biggest challenges.


InfiniBand (IB) Wide Area Networking
SC 05
Seattle
HD RAVE
Video Storage
Abilene & HOPI
IPv6
NRL
DC
High-Speed Wide-Area Secure Peer-to-peer
Distributed Computing Functionality needed by
DoD/IC, NASA and DOE
– SuperComputers (as if) on your desktop

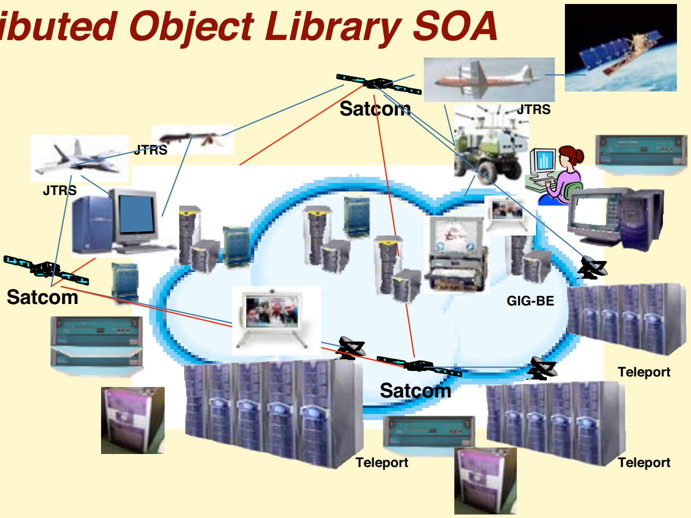

100G
10G
1G
100M
1M
SuperCs
Sensors
Teleports
JTRS, Wireless, Mobile, etc …
Network-Centric Bandwidth Amplification
Large Numbers of Lite-Clients

OPERATIONAL REQUIREMENTS

- Global access to the *"right data"* instantly
- Same "right data" everywhere (cache-coherent, synchronized)
- Flexible access for global ***REACHBACK***
- Intuitive access to Large Data Sets (petabytes to exabytes in magnitude)
- Composable remote visualization of large data
- ***TRACEBACK*** for change analysis on an unprecedented scale for signature development, pattern recognition, targeting, forensics, etc.
- ***"Global Information Grid"*** net-centric extension to warfighters deployed or afloat

JCTD: Interactive Distributed Object Library SOA

❖Virtual network of Active Information Producers & Consumers
… i.e., Grid core w/ P2P edges
❖Vertical fusion - aggregation, delegation
… i.e., level of detail
❖ Horizontal fusion - peer group metadata search & discovery
… e.g., DoD Discovery Metadata Standard
❖Agile data type support for spatiotemporal indexing
❖Pluggable transport architecture including IPV6, native ATM & hardware QoS, DWDM
❖Intelligent caching hierarchy for multi-terabyte/petabyte datasets (BIG DATA …)



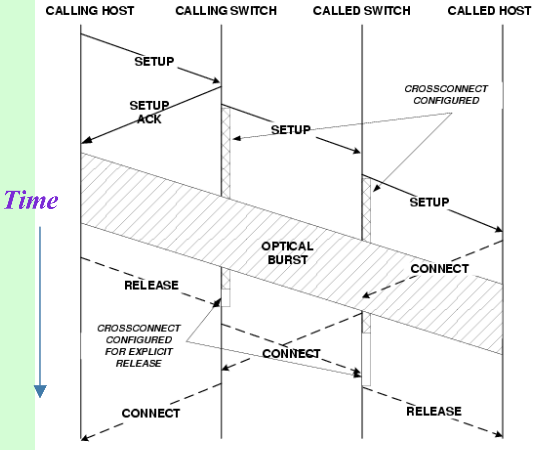
Distributed Database Backend

❖ Immersive Zoomable User Interface (ZUI)
❖ Filter and layer definition, selection, and presentation support
❖ Flexible, intuitive manipulation
❖ Platform support ranging from PDA to workstation to distributed grid to HPCS supercomputer
… High performance: SGI InfiniteReality & UltimateVision systems … well defined API
… Ubiquitous: Desktop PC/Mac/Linux, open source
… Pervasive: iPAQ handheld

Visualization Front End

Scalable *Optical Burst Switching … JIT Signaling*

- *No round-trip delay (for 2- or 3-way handshake) required prior to data burst*
- *Out-of-band signaling message precedes data burst*
- *A signaling message's lead time over its data burst shrinks as both propagate through network*
- *Switch resources held only for the duration of burst; no light path required*
- *JIT simplicity - smaller, lighter hardware processing modules*
- *SIP control plane initiated*



Single Burst Example

- *Significant improvement in throughput and determinism vs TCP/IP/GMPLS*
- *Out-of-band JIT signaling increases communications security & reliability*

Several KEY Observations . . .

- Large Scale National ***TERABIT*** Core Optical Testbed req'd
  - ***Large data applications will NOT adopt toy infrastructures***
  - ***Simulation/Emulation is NOT a total substitute for REAL WORLD test***
- Infrastructure research is highly interdisciplinary
  - ***HPCC, e-Science: Medical, HEP, Visualization, Web, Voice, Cellular, …***
- Infrastructure goals not well understood by Gov't Agencies
- Current Infrastructure Research Funding is Insubstantial
- *Japan, Canada, parts Europe, China* already underway
- Infrastructure Research requires:
  - ***Long Term Objectives/Investments***
  - ***Scalable, transparent core abstractions***
  - ***Commitment to Long Term Vision***
  - ***Flexibility to test (break): EARLY and OFTEN***

SUMMARY . . . *Bandwidth with Lowest Latency = Professionalism!*

A challenge for ***Net-centric Architectures*** is to provide an ***integrated, lowest-latency DISTRIBUTED, FEDERATED INFRASTRUCTURE*** core that supports moving bits to large data flows globally with QoS & IA E2E

Next-Gen optical and IP services require a more flexible transport layer

Powerful set of ***new technical capabilities*** are essential … Worldwide ***GLIF*** lambda's and ***Infiniband*** I/O to meet the challenges of transporting large data flows with low latency through interconnected service grids: ***RAIN*** grids for scaled online large data repositories; computational ***MPI***-based grids; ***VIZ***ualization grids; ***P2P*** gaming; ***VoIP*** and ***3G*** cellular

Need to advance leading edge terabit low-latency flow research by establishing and maintaining a nationwide advanced network infrastructure to interoperate with GLIF

"Expose TERABIT interfaces early and often!"

DEPARTMENT OF THE NAVY ★ NAVAL RESEARCH LABORATORY ★

Thank You

Center for Computational Science

of the Naval Research Laboratory