INFINIBAND

Prototyping Future Terabit Networks

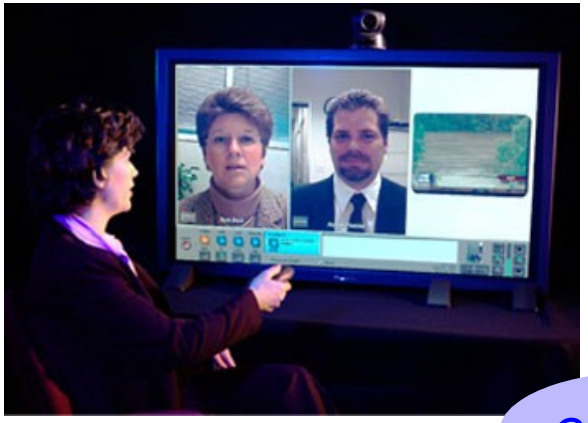
Naval Research Laboratory

A Terabit Challenge . . .

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

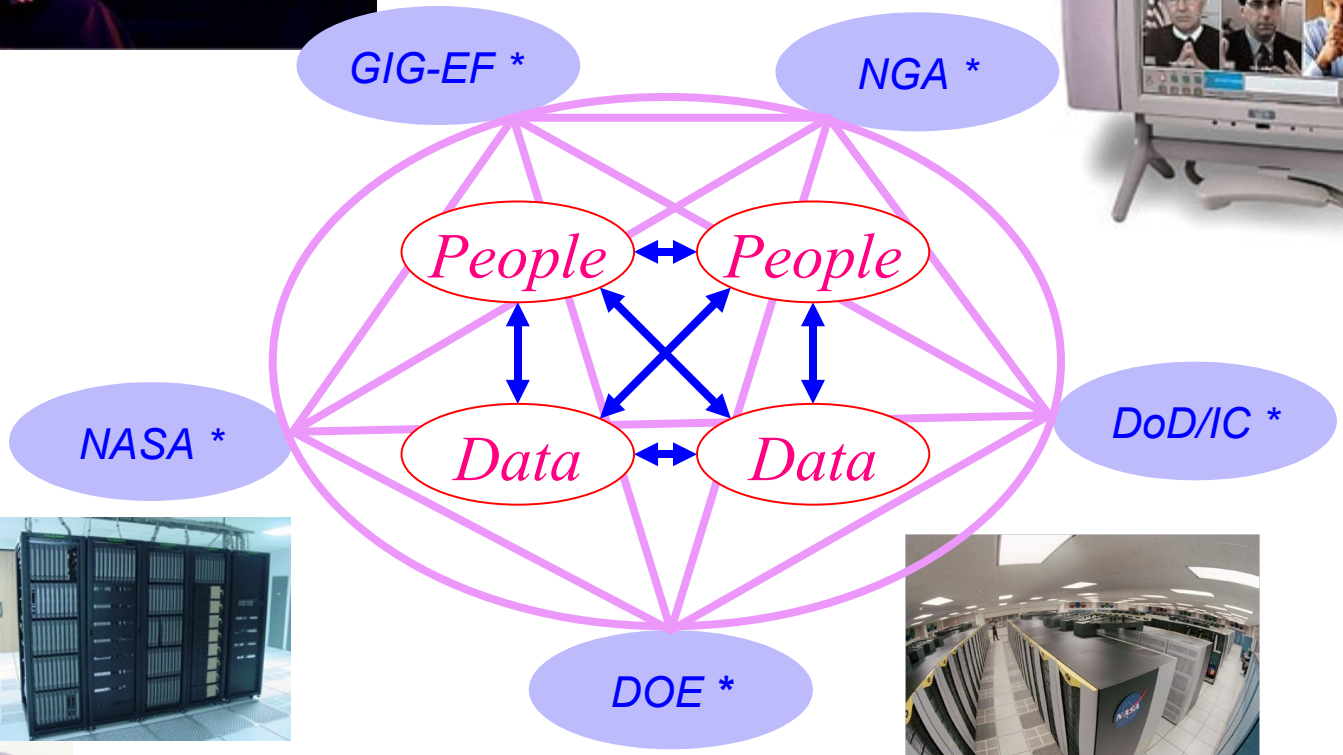
- *Integrate **federated, distributed** computational grids, realtime sensors, and digital historical information*
- *Scale to support **exponentially** increasing data archives*
- *Privacy, authenticity and security demands: **InfoAssured***
- *Affordable … highly available … **E2E QoS/QoP** flows*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

"Expose interfaces early and often …"



An Enterprise View …

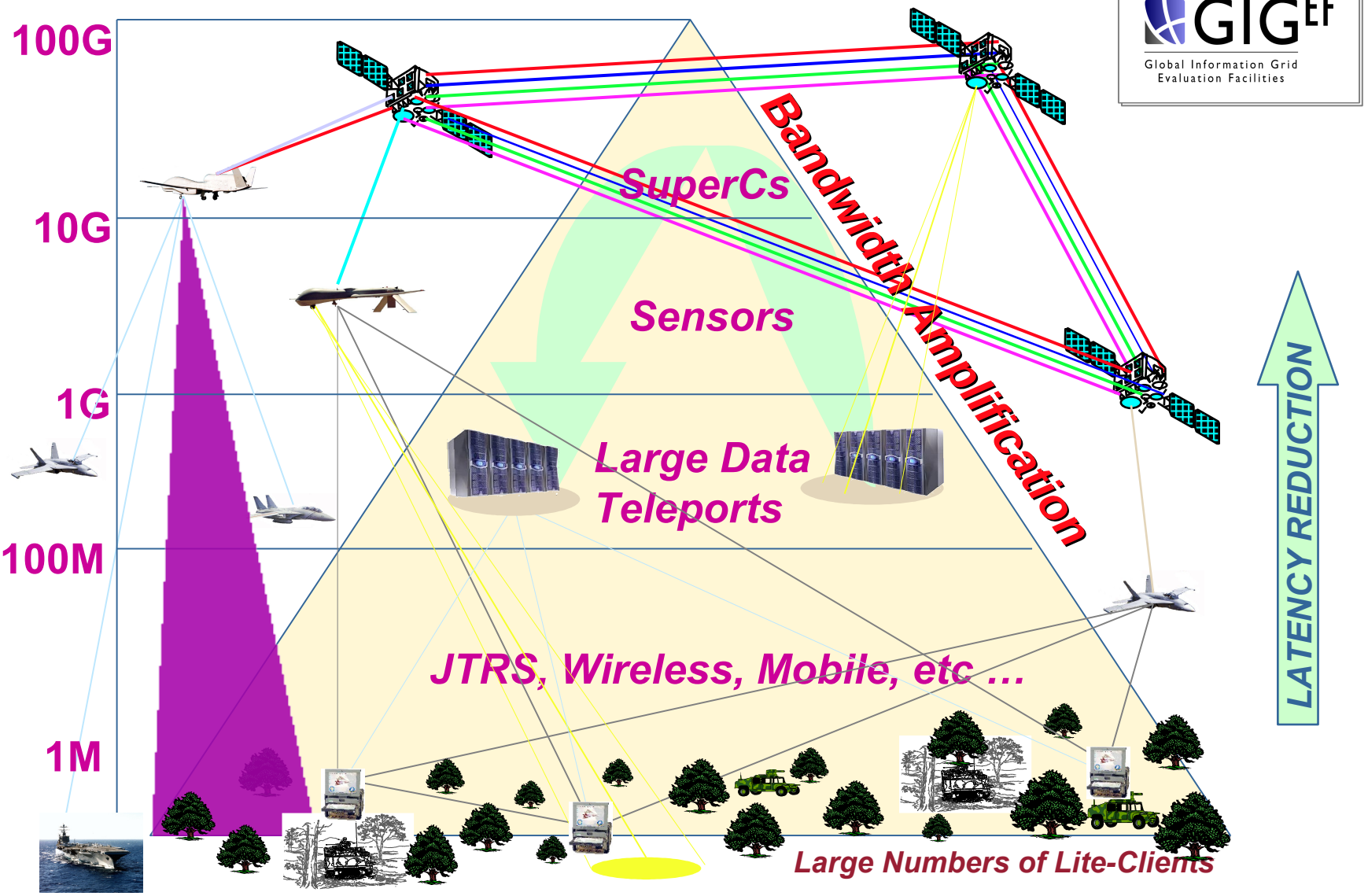
* Hypothetical sites



"DATA CONFERENCING"

… multiple sites, people, P2P seamlessly interacting!

A Net-centric Architecture . . .



"… a single packet triggers High Bandwidth Flows …"

big *fast* *"terabytes/hour"* data problem …

… efficiently interface high performance optical networks directly to

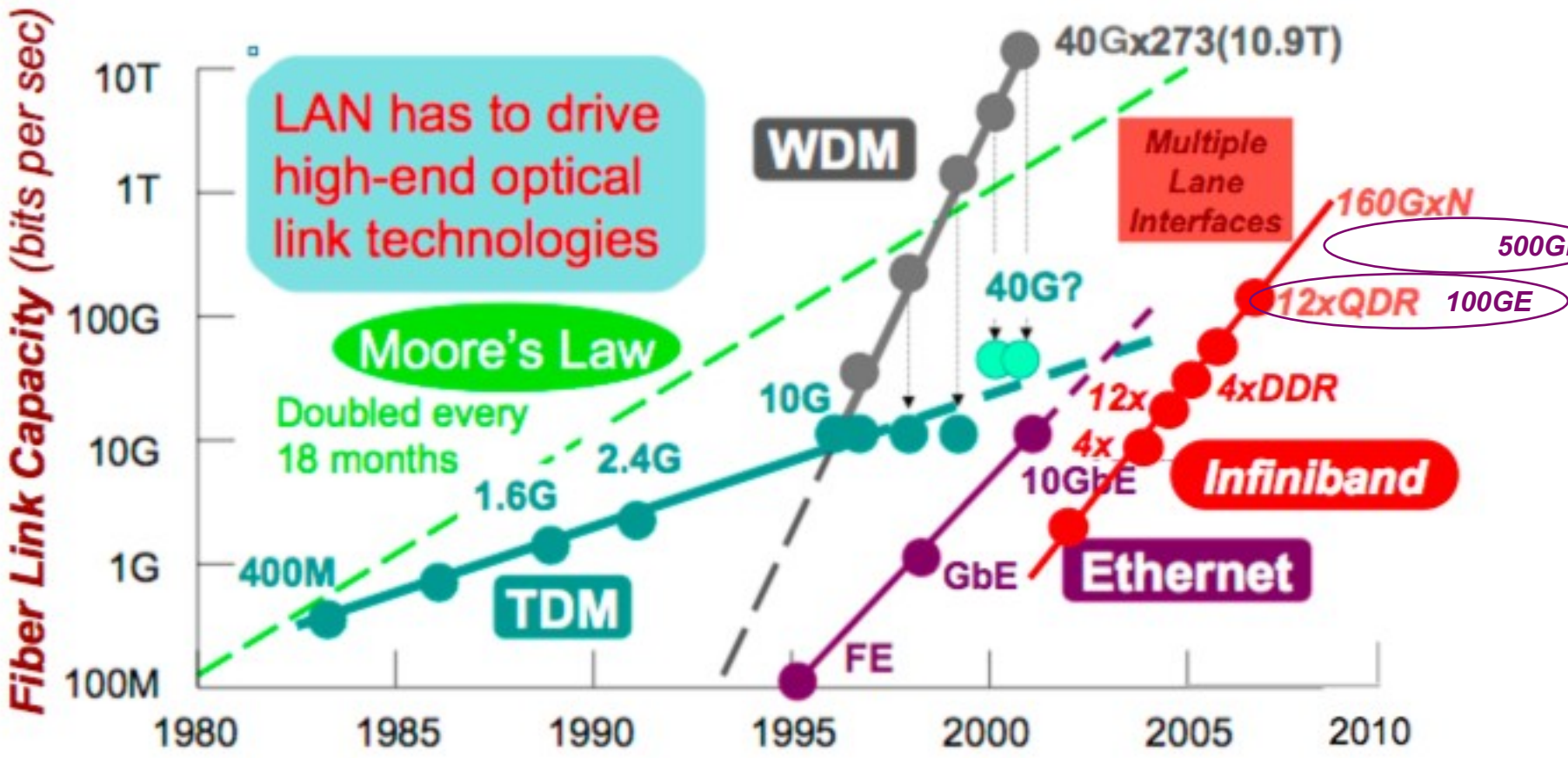
- Supercomputers
- Grid Clusters
- Visualization, SuperHDTV
- HR Motion Imagery
- TSAT Tactical Comms
- GIS Imagery/Weather/Oceans
- 2D/3D workstations
- Online Digital Asset Archives
- Hyperspectral …40K x 40K
- Virtualized Ground Station

- 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*

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

-Multi-PetaByte online distributed, federated archives

Optical Link Performance, per Laser



Ref: O. Ishida, NTT, "Toward Terabit LAN/WAN" Panel, iGrid 2005

Optical Technology Forecast . . .

Yesterday

- *Dispersion Compensation Fiber (DCF) hardwired*
- *No control plan, bandwidth management for lambda services*
- *Static point-to-point optical links, rings and OADMs*
- *Partitioned Access / Metro / LH / ULH optical transport solutions*
- *No layer awareness*

Today

- *Improved economics via large scale photonic integration*
- *Electronic Dispersion Compensation (EDC)*
- *End-to-end GMPLS control plane*
- *Reconfigurable OADMs for wavelength interchange*
- *Dynamic meshed optical nets with flexible Bandwidth Management for wavelength services*
- *Integrated Access/Metro/LH/ULH optical transport solutions*

Within 5 Years

- *≥ 1.6 Tbps very large scale photonic integration*
- *≥ 40 Gbps individual channels*
- *≥ 100 Gbps concatenated Channels (Nx100 GbE)*
- *≥ 10 Tbps optical line systems*
- *Multi-domain (L3/L2/L1) IP/DWDM integrated optical networks*

Within 10 Years

- ≥ 1.0 Tbps single optical flows … switched lambdas E2E

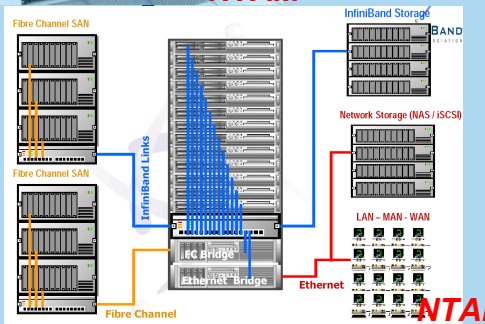
Reference: ONT2 Workshop, NASA Ames, September 2005 … D. Perkins,Infinera & D. Blumenthal, UCSB

Infiniband: A Single Wire Solution

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

Processing

NTAM



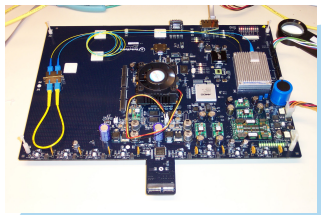
NTAM

Storage

InfiniBand

NTAM

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



Campus

NTAM Firewalls

WAN

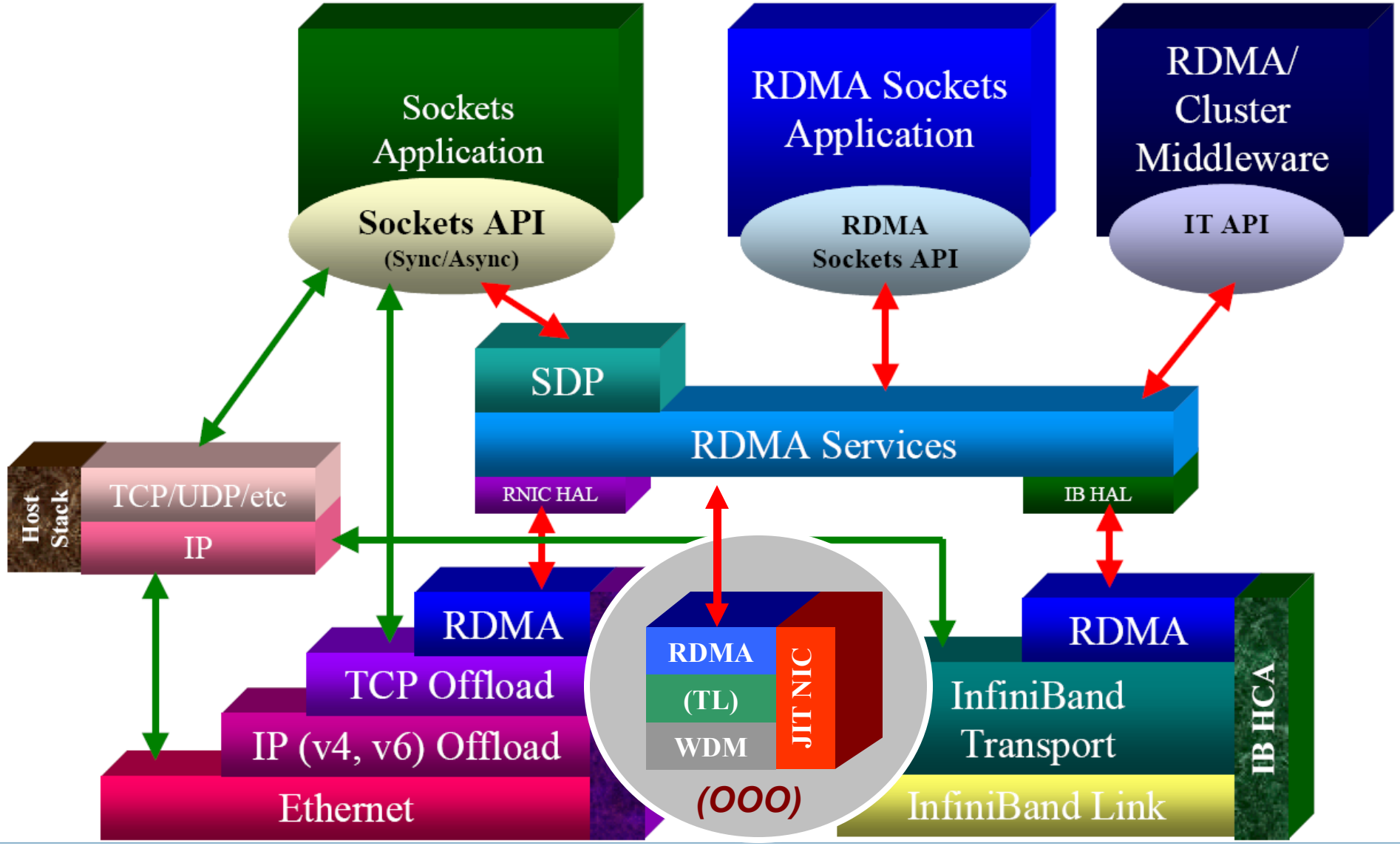
Communications

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

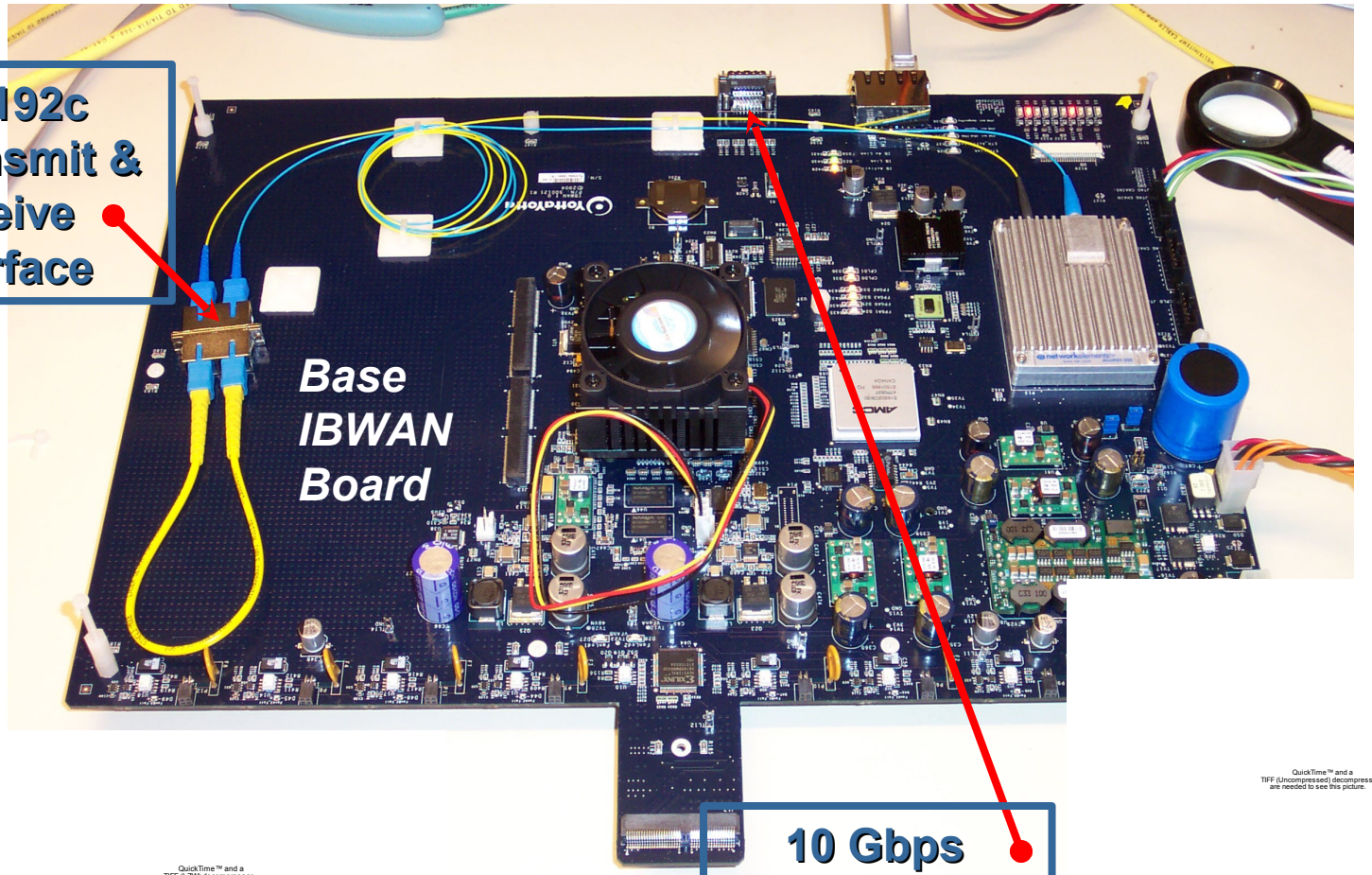
According to Intel
http://www.intel.com/technology/infiniband/whatis.htm

RDMA Infrastructure:
Solution Components

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



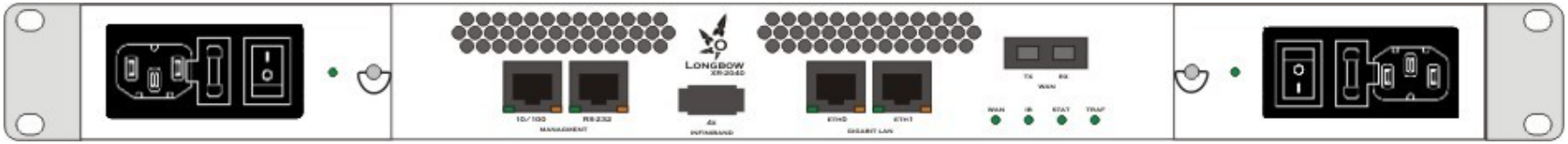
IBWAN: Functional Prototype …



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 (Q1'06)
- Applications software being developed to facilitate deployment of wide area *switched wavelength* IB data streaming technology



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 . . . CY2005/6

Point-to-point:

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

Targeted: 3-way multicast

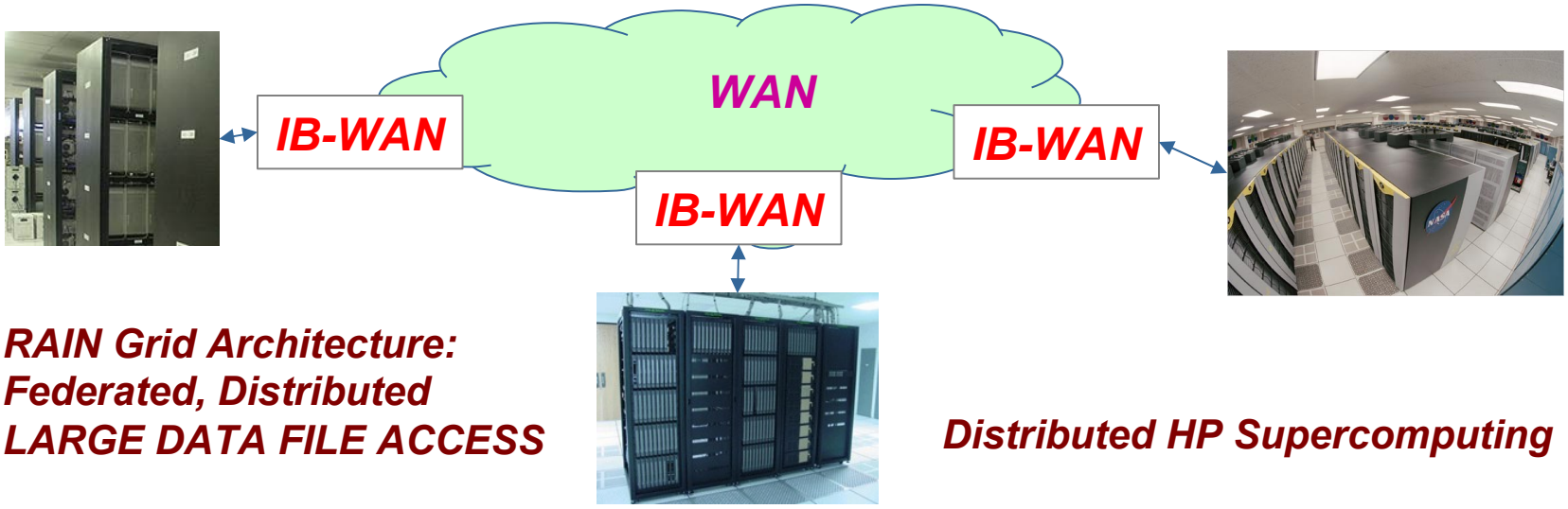
- *ATM with QOS (OC-192c or OC-48c)*
- *IPv6 POS (OC-192c or OC-48c or 10 GigE)*
- *GMPLS (preset)/ JIT (OBS research)*
- *SMPTE 292m (4:2:2 & 4:4:4) 720p/1080p*

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**

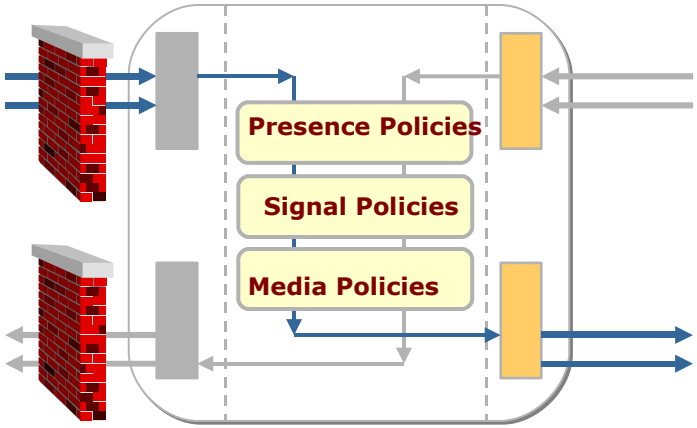
A request-response protocol that closely resembles HTTP & SMTP

- **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!"
… Vint Cerf

1. Internet Communications Using SIP – Sinnreich, Johnston, John Wiley & Sons, 2003

***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 Real-Time Flow Monitor

- 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

Performance

- OC-192c, 10GB Ethernet, OC-48c, OC-12c, 100/10 MB Ethernet, SLIP
- *Ongoing research to scaling to OC-768c*
- POS, ATM, Ethernet, FDDI, SLIP, PPP
- > 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, all protocols, unencrypted or encrypted, at all layers and for all levels of encapsulation !

Network Service Functional Assurance

- *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

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

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

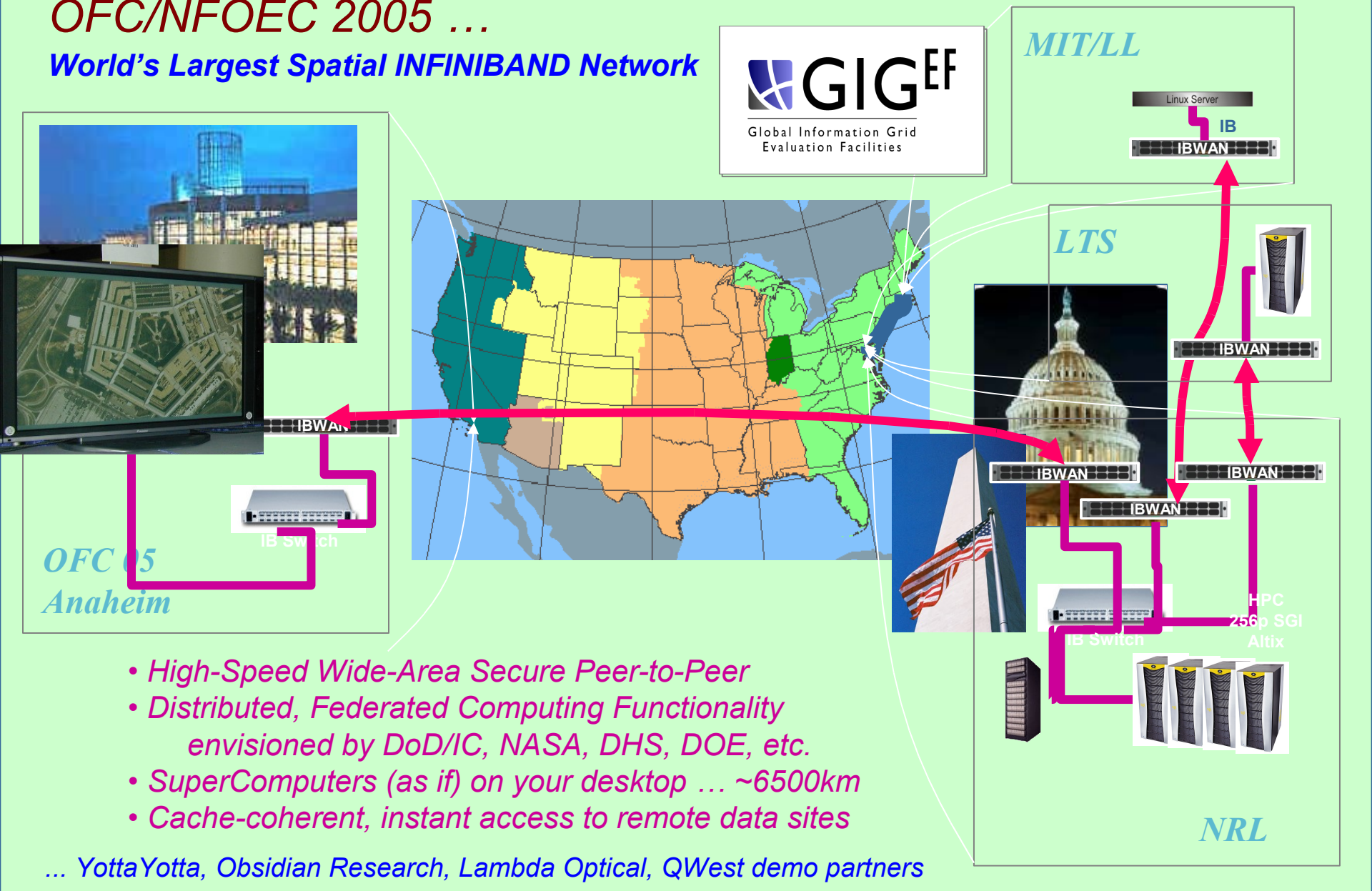
SPECIAL TOPICS	Quantum Key Distribution (QKD), Dynamic PMD Comp, Peering/Multicast, Parallel Optics, OOO(2R) Optical Regeneration, . . .

InfiniBand *Wide Area Networking*

OFC/NFOEC 2005 …

World's Largest Spatial INFINIBAND Network

GIG-EF
Global Information Grid
Evaluation Facilities

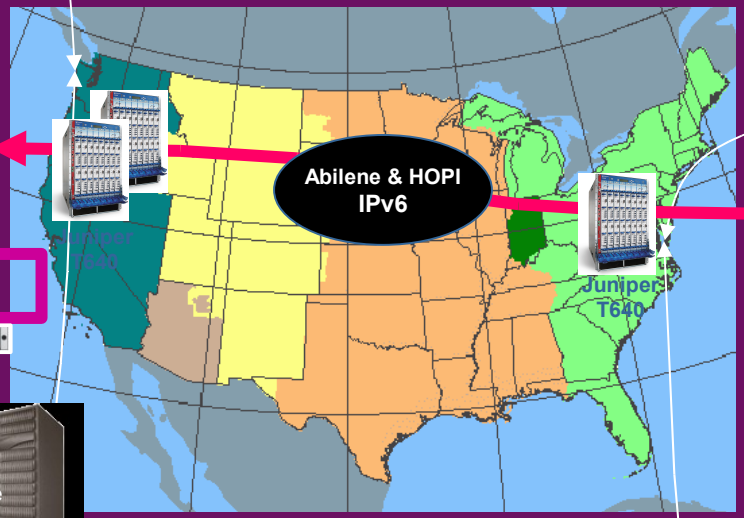


- *High-Speed Wide-Area Secure Peer-to-Peer*
- *Distributed, Federated Computing Functionality envisioned by DoD/IC, NASA, DHS, DOE, etc.*
- *SuperComputers (as if) on your desktop … ~6500km*
- *Cache-coherent, instant access to remote data sites*

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

InfiniBand (IB) Wide Area Networking …

SC2005



IBWAN

IB

Voltaire ISR9024

HD RAVE
Video retrieval

YottaYotta GSX3000

SGI Prism

Voltaire ISR9024

IB storage

OPEN IB ALLIANCE

SC 05 Seattle

IBWAN

Voltaire ISR9024

YottaYotta GSX3000

NRL DC

Voltaire ISR9096

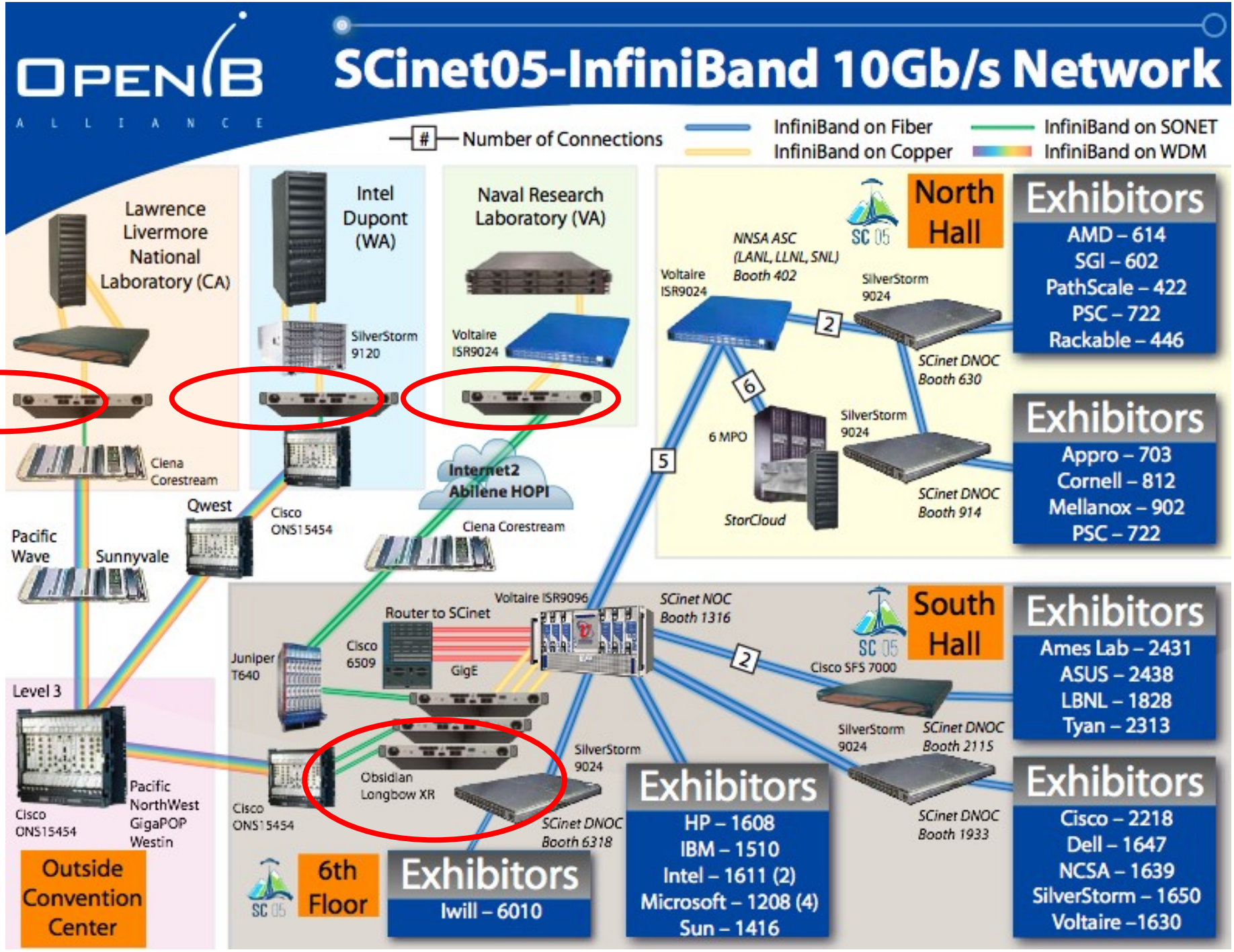
6p SGI Altix

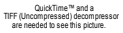
SGI Prism

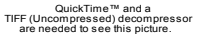
GIGEF

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

– SuperComputers as if on your desktop …


OPEN IB ALLIANCE
SCinet05-InfiniBand 10Gb/s Network
Number of Connections
InfiniBand on Fiber
InfiniBand on Copper
InfiniBand on SONET
InfiniBand on WDM
Lawrence Livermore National Laboratory (CA)
Intel Dupont (WA)
Naval Research Laboratory (VA)
SilverStorm 9120
Voltaire ISR9024
Ciena Corestream
Internet2 Abilene HOPI
Qwest
Cisco ONS15454
Pacific Wave
Sunnyvale
Level 3
Cisco ONS15454
Pacific NorthWest GigaPOP Westin
Outside Convention Center
North Hall
NNSA ASC (LANL, LLNL, SNL) Booth 402
Voltaire ISR9024
SilverStorm 9024
SCinet DNOC Booth 630
6 MPO
StorCloud
SilverStorm 9024
SCinet DNOC Booth 914
Exhibitors
AMD – 614
SGI – 602
PathScale – 422
PSC – 722
Rackable – 446
Exhibitors
Appro – 703
Cornell – 812
Mellanox – 902
PSC – 722
South Hall
Voltaire ISR9096
SCinet NOC Booth 1316
Router to SCinet
Cisco 6509
GigE
Juniper T640
Cisco ONS15454
Obsidian Longbow XR
SilverStorm 9024
SCinet DNOC Booth 6318
Cisco SFS 7000
SilverStorm 9024
SCinet DNOC Booth 2115
SCinet DNOC Booth 1933
Exhibitors
Ames Lab – 2431
ASUS – 2438
LBNL – 1828
Tyan – 2313
Exhibitors
Cisco – 2218
Dell – 1647
NCSA – 1639
SilverStorm – 1650
Voltaire –1630
Exhibitors
HP – 1608
IBM – 1510
Intel – 1611 (2)
Microsoft – 1208 (4)
Sun – 1416
6th Floor
Exhibitors
Iwill – 6010





Nov 15, 2005 10:52

Joint Industry and Government Initiative to Demonstrate Long Distance InfiniBand(R) at SC05

SEATTLE --(Business Wire)-- Nov. 15, 2005 Cisco Systems, Intel Corporation, Lawrence Livermore National Laboratory, Microsoft, Naval Research Laboratory, Obsidian Research, the OpenIB Alliance and Qwest Communications today announced they are demonstrating extended computing resources using InfiniBand(R) technology as part of SCinet at SC05. Sponsored by the IEEE and ACM, SC05 is the premier international conference on high performance computing, networking and storage.

These leading telecommunications and technology organizations are jointly providing the equipment, software and applications for industry and research partners to demonstrate the value of high-performance, low-latency direct access networking at 10 Gigabits per second over a long distance infrastructure. InfiniBand is a high performance, switched fabric interconnect standard for servers and OpenIB is an industry Alliance that supplies open source InfiniBand software. Both are quickly becoming the preferred standard in high performance computing, grid and enterprise data centers.

• • • • •

"Microsoft's keynote demonstration at Supercomputing 2005 showcased the improved productivity for scientists made possible by seamless access from the workstation to structured data stores, personal desk-side clusters for interactive analysis and large heterogeneous pools of computing resources for detailed studies," said Kyril Faenov, director of high performance computing, Microsoft Corp. "The high-bandwidth connectivity to Intel's Dupont location allowed us to seamlessly incorporate a 256-core Intel(R) Xeon(TM) cluster running Windows Compute Cluster Server 2003 to the mix of computing resources."

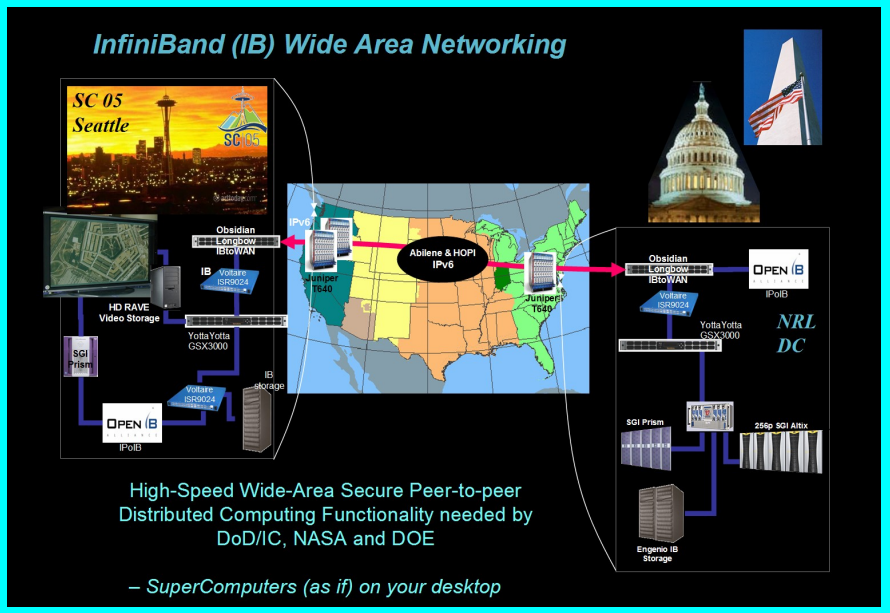
SCALING THE GLOBAL INFORMATION GRID

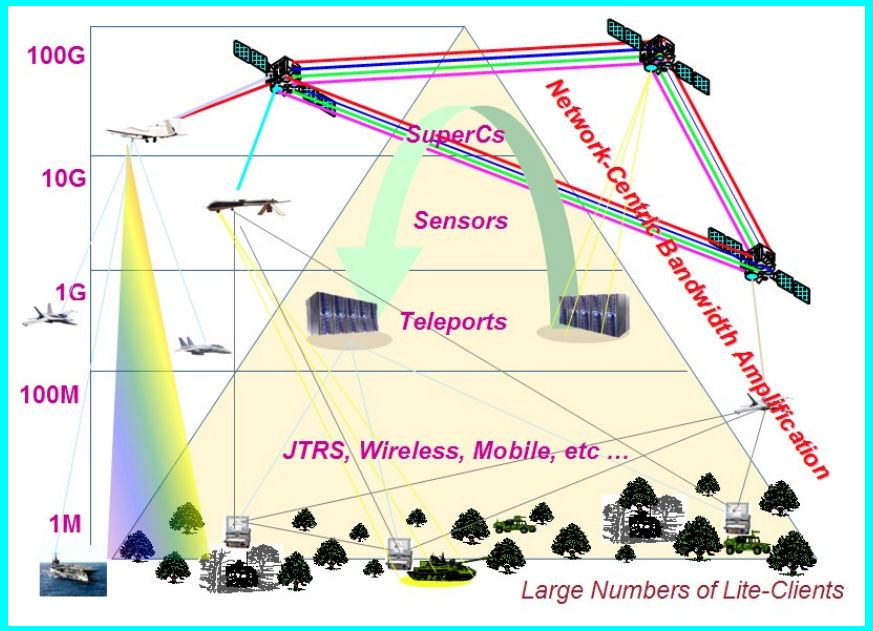
Naval Research Laboratory

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.



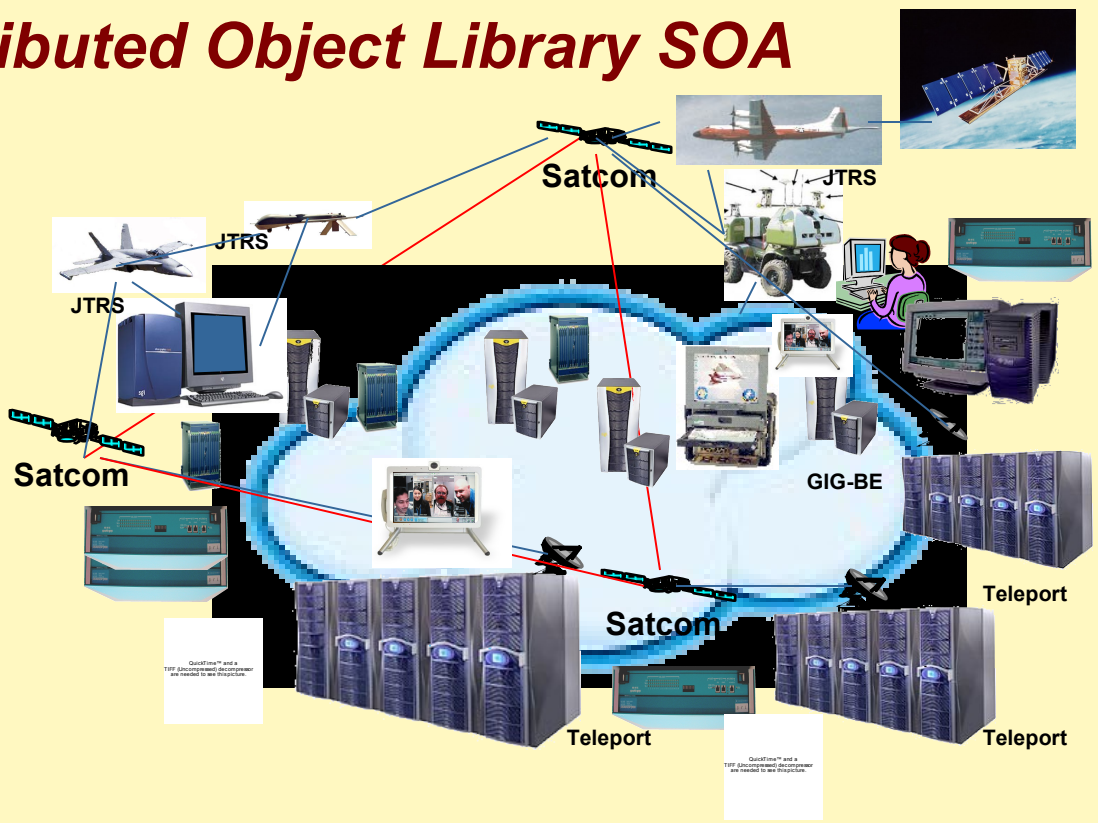


OPERATIONAL REQUIREMENTS

- 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

"Working for Terabits"

Thank You

Center for Computational Science
of the Naval Research Laboratory