DEPARTMENT OF THE NAVY ★ NAVAL RESEARCH LABORATORY ★

# Terabit Technology Vision

## iGrid 2005

28 September 2005

Henry D. Dardy
Naval Research Laboratory
Washington, D.C. 20375

# Terabit Challenge . . .

Build a Terabit global Integrated Information Infrastructure to improve the ability to Rapidly Produce Knowledge from the Best Information available.

- Integrate distributed, federated computational resources, realtime sensors, and historical information
- Scalable to support exponentially increasing data
- Privacy, authenticity and security demands
- Affordable … highly available … E2E QoS/QoP
- Legacy and rapidly evolving technology issues
- Performance, NetOps, Information Assurance tools

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

… efficiently interface high performance optical networks directly to

- *Supercomputers*
- *Grid Clusters*
- *Visualization*
- *Motion Imagery*
- *Imagery/Weather/Oceans*
- *2D/3D workstations*
- *Digital Assets Archives*
- *Hyperspectral …40K x 40K*

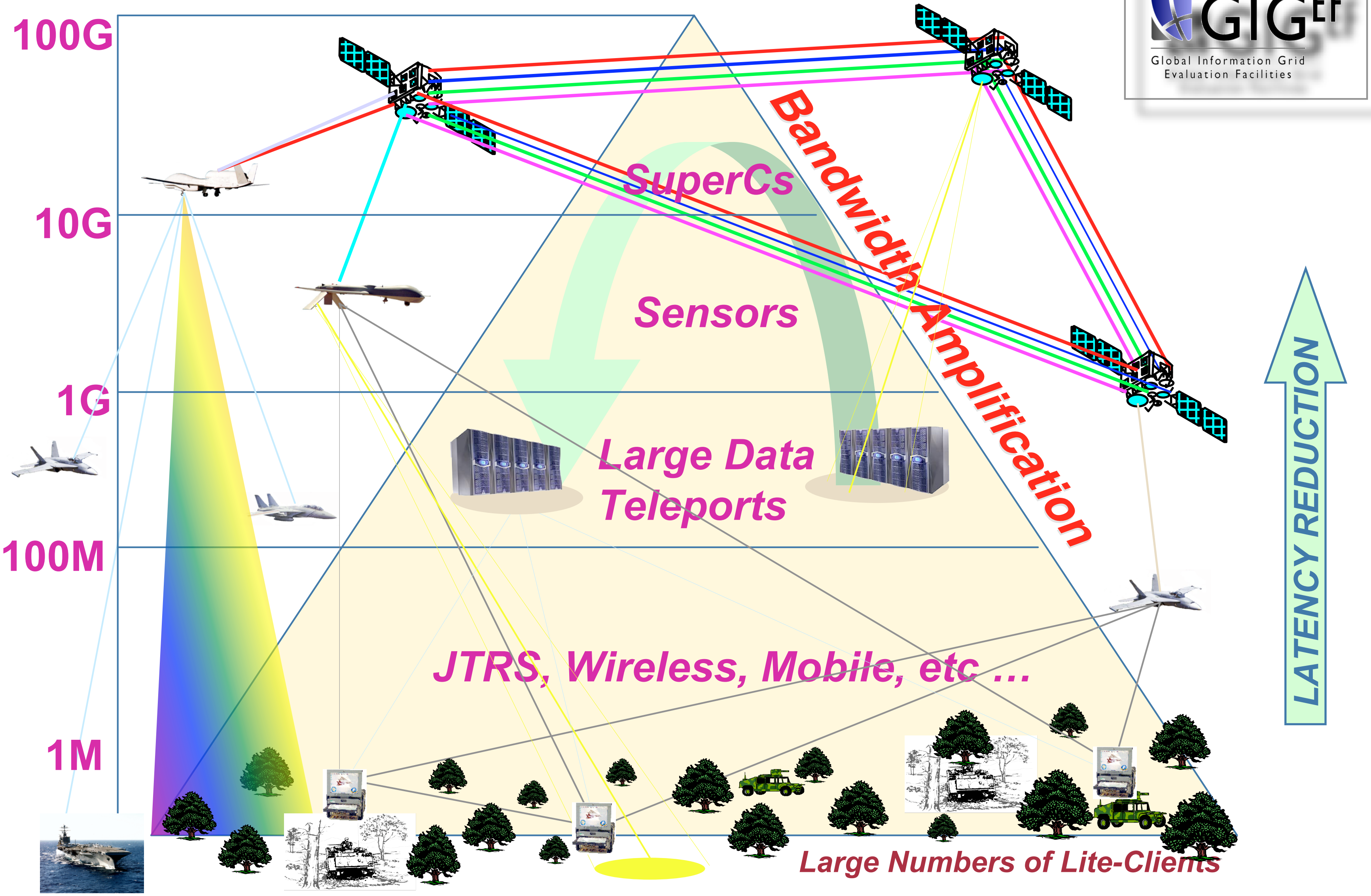
- Interfaces scale as optical networks scale
- Interface programming model and semantics familiar and friendly
- Minimum of equipment required for each lambda connection
- WAN transport protocol semantics abstracted from applications
- Sustained performance across the WAN approaches full wire speed

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

*-Multi-PetaByte online distributed, federated archives*

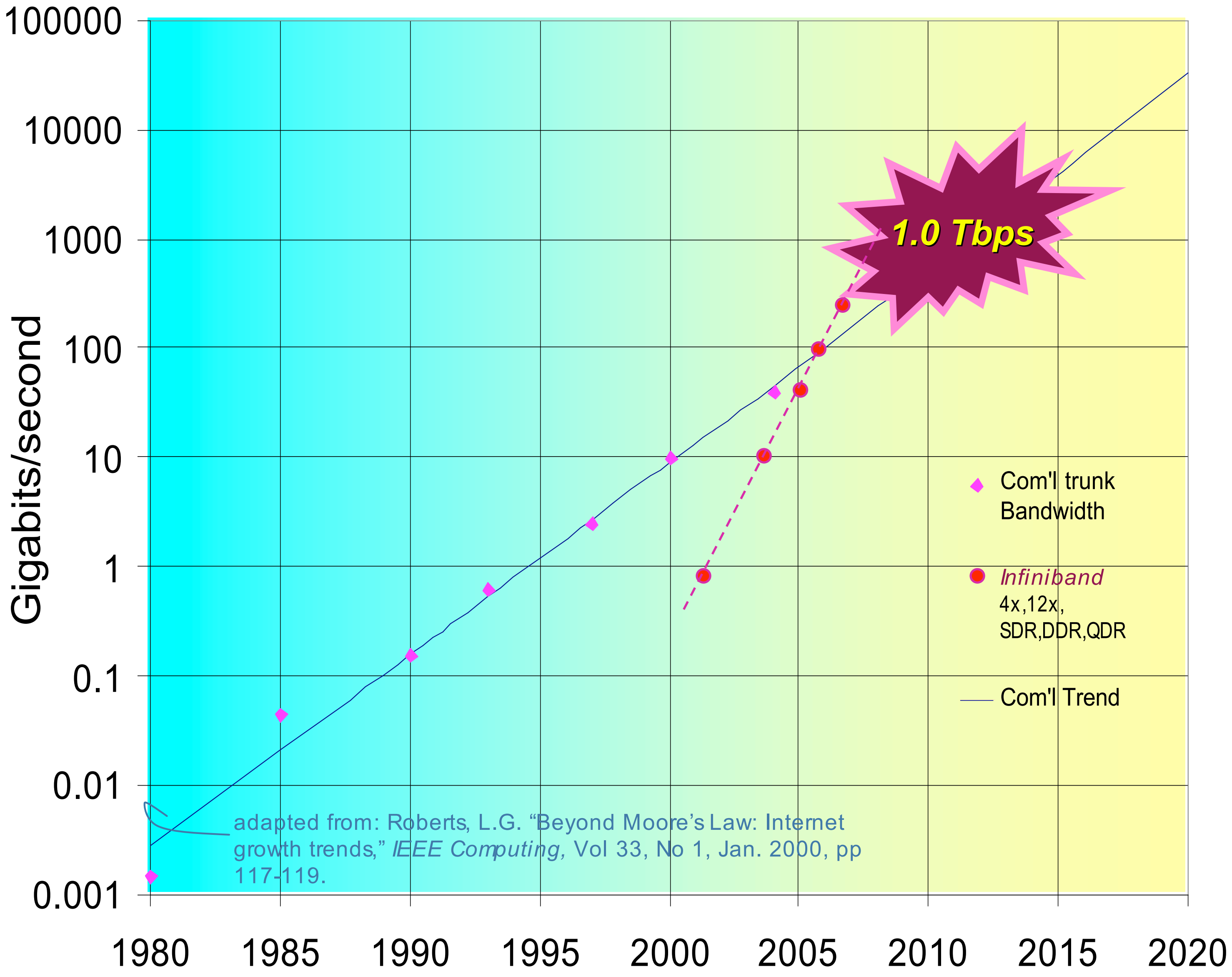
# Net-centric Architecture . . .

GIG^EF
Global Information Grid
Evaluation Facilities



*"… a single packet triggers High Bandwidth Flows …"*

# Network Growth Trend . . .

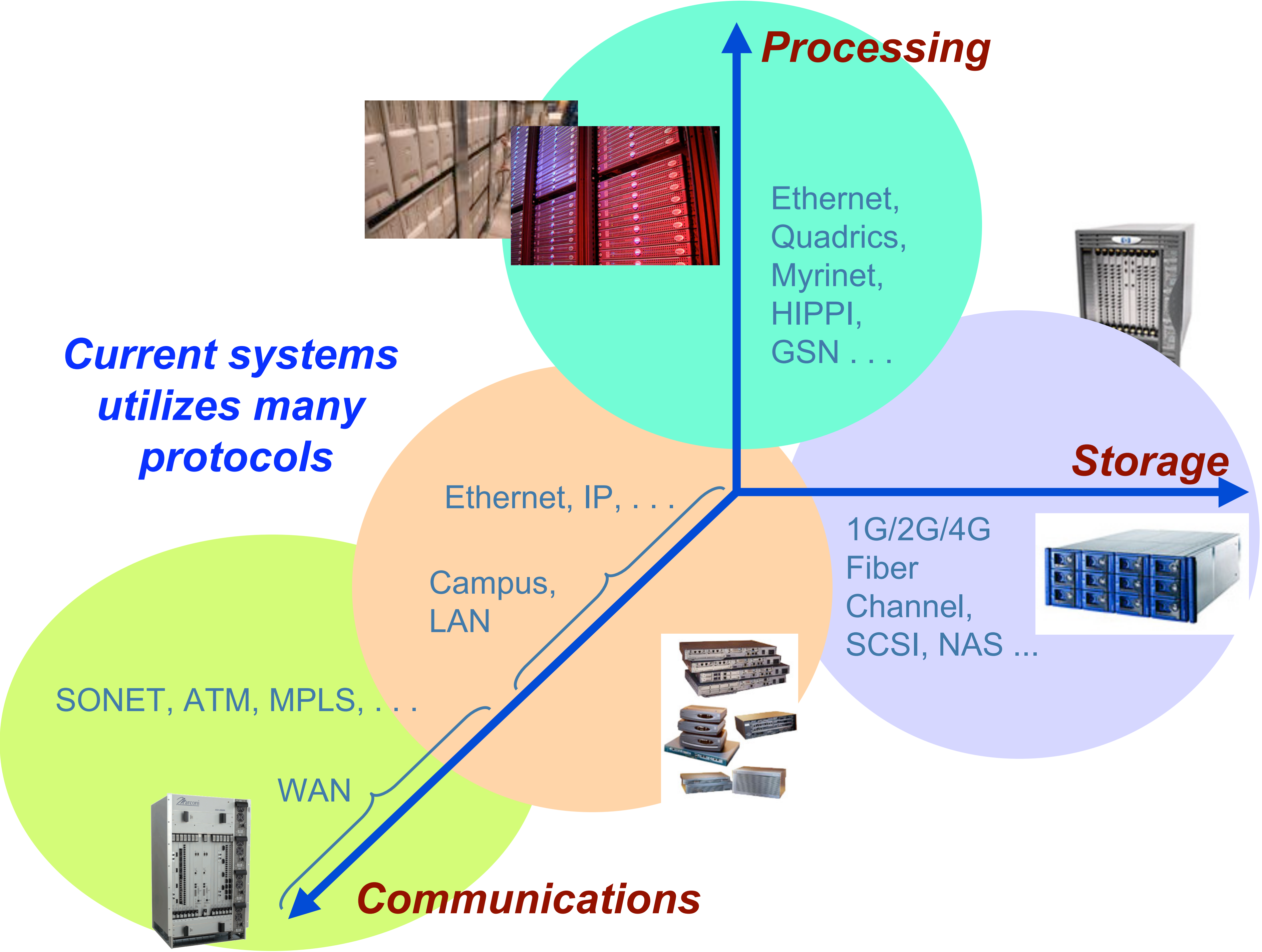


adapted from: Roberts, L.G. “Beyond Moore’s Law: Internet growth trends,” *IEEE Computing,* Vol 33, No 1, Jan. 2000, pp 117-119.

# *How do we get there . . . ???*

- *Terabit end-to-end **low-latency** streaming*
- *Data Conferencing: P2P, D2D, P2D, D2P*
- *Virtualized Enterprise: RSOCs, Distributed Ground Sites, Federated assets, etc.*
- *Seamless space, terrestrial, mobile, wireless access to Large Data Archives*
- ***Traceback** and **Reachback** into data stores*
- *High resolution motion imagery, starred imagery, hyperspectral imagery, etc.*
- *Lambda(s)-to-the-Edge…**streams bypass***

***Current systems utilizes many protocols***



***Processing***

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

***Storage***

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

***Communications***

Ethernet, IP, . . .

Campus,
LAN

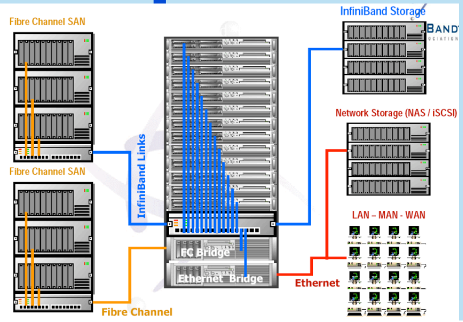
SONET, ATM, MPLS, . . .

WAN

# InfiniBand Integrates High Performance Information Systems!!

Cam

Processing

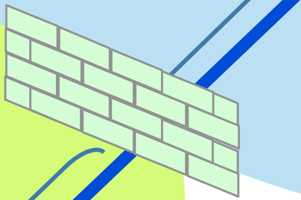


http://www.infinibandta.org/events/past/it_roadshow/overview.pdf

Storage

InfiniBand

Campus



Router Filter
Firewalls

tcp

IPV6, ATM, MPLS, …

WAN

Communications

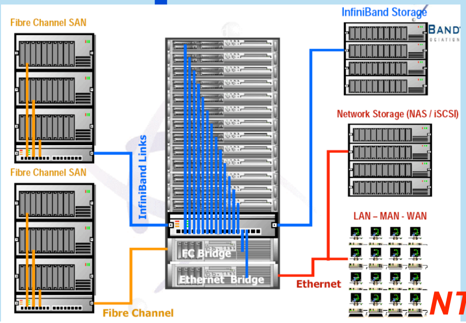
- Greater performance,
- Lower latency,
- Easier and faster sharing of data,
- Built in security and
- Quality of Service,
- Improved usability
- Reliability
- Scalability

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

## InfiniBand to WAN Gateway adds the secure WAN to the integrated InfiniBand domain.

Processing

NTAM



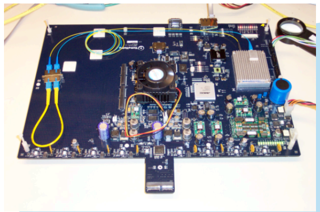
NTAM

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

Storage

InfiniBand

NTAM



Campus

NTAM
Firewalls

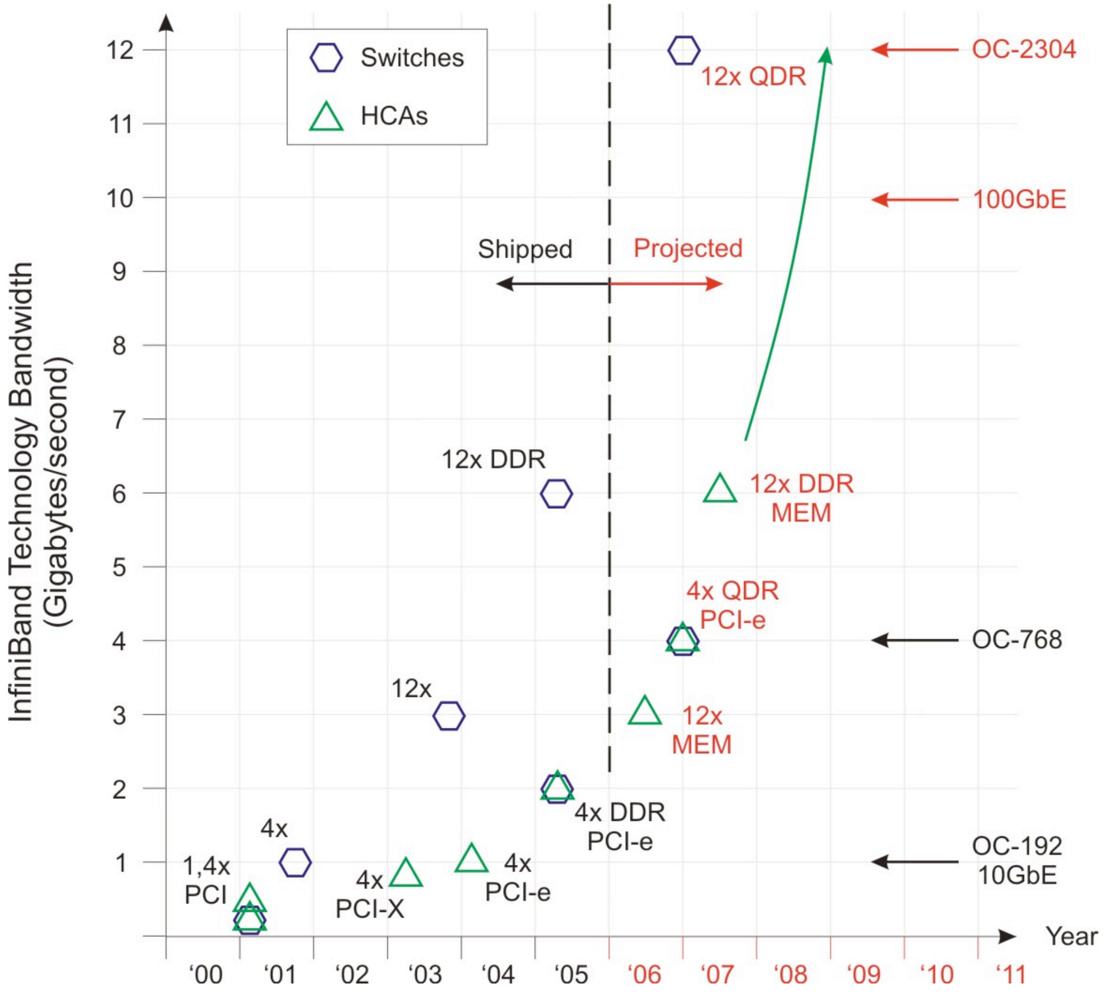
WAN

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*

# Today 10Gbps INFINIBAND … and beyond !



**InfiniBand will outpace WAN technology (since HPC is the driver…)**

**Channels may be extrapolated in both individual bandwidth and number**

**Links may be aggregated in switches – allows for scaling beyond individual node memory bandwidths**

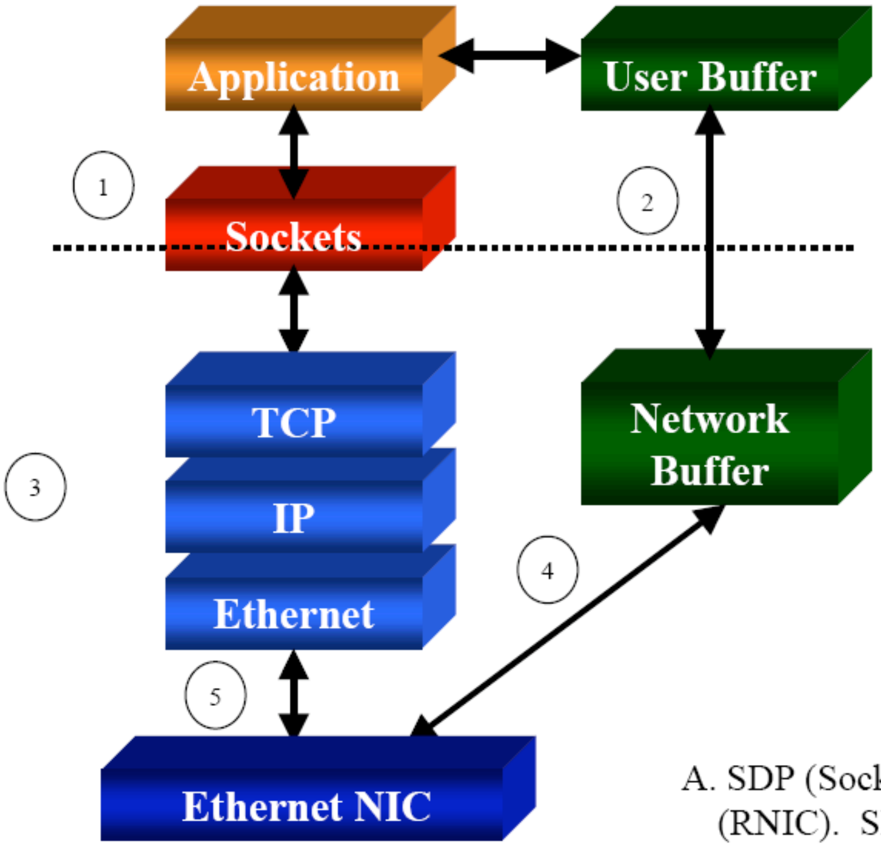
**Continued bandwidth scale up feasible thanks to the bypassing of:**

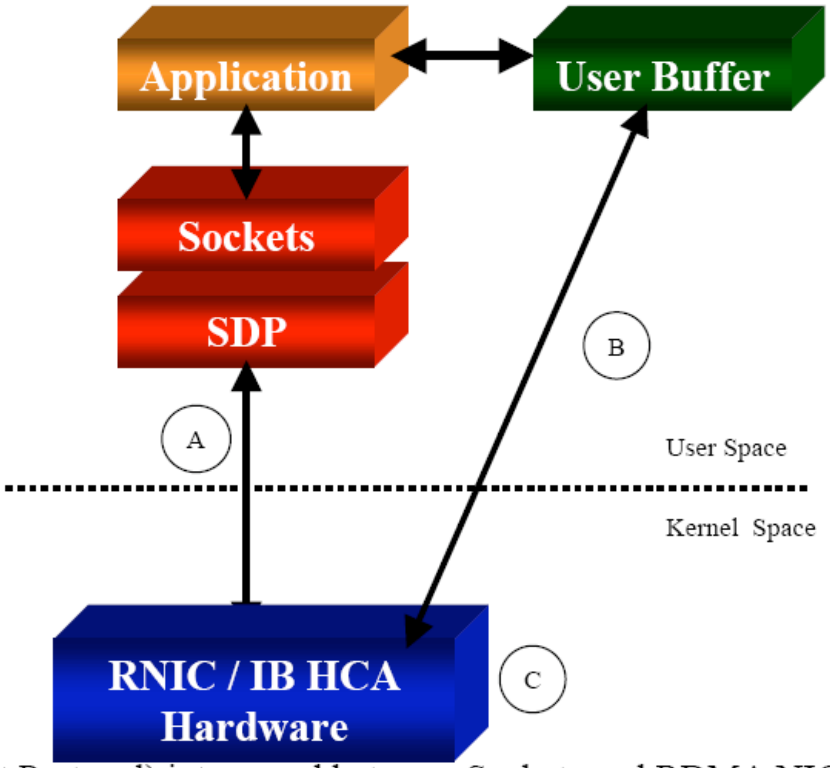
- **OS / drivers (RDMA)**
- **IO bus replacement**
- **CPU core: direct to memory**

**Prospect of continuity for applications codes and equipment infrastructure**

# Existing Architecture ➡ RDMA Architecture

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





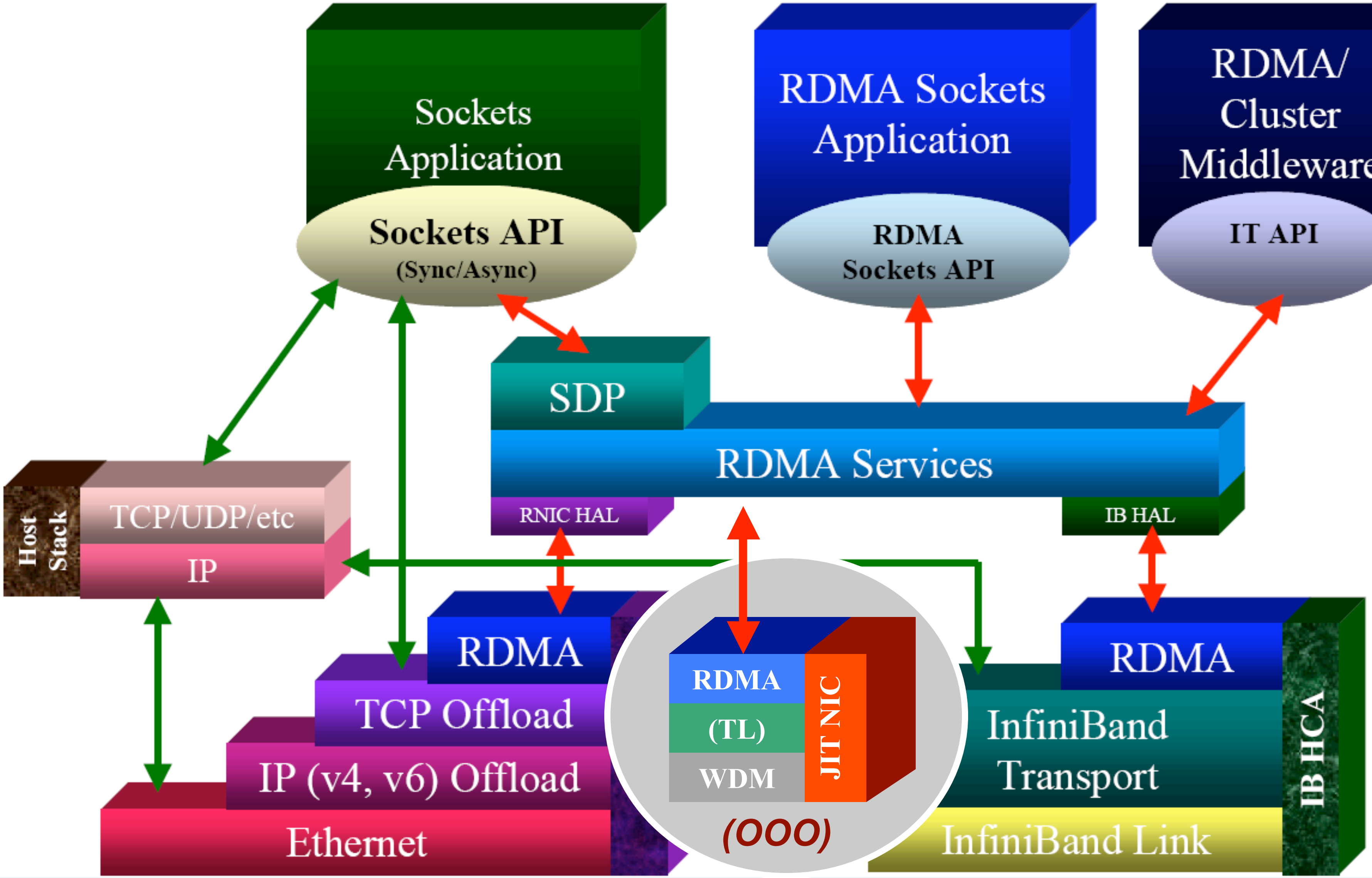
A. SDP (Sockets Direct Protocol) interposed between Sockets and RDMA NIC (RNIC). SDP enables SOCK_STREAM applications to transparently operate over RNIC. SDP interacts with the RNIC directly to process application and SDP “middleware" message exchanges. Enables OS Bypass.

B. Direct DMA to / from user buffer. No interrupts are required as completion processing is performed within SDP layer.

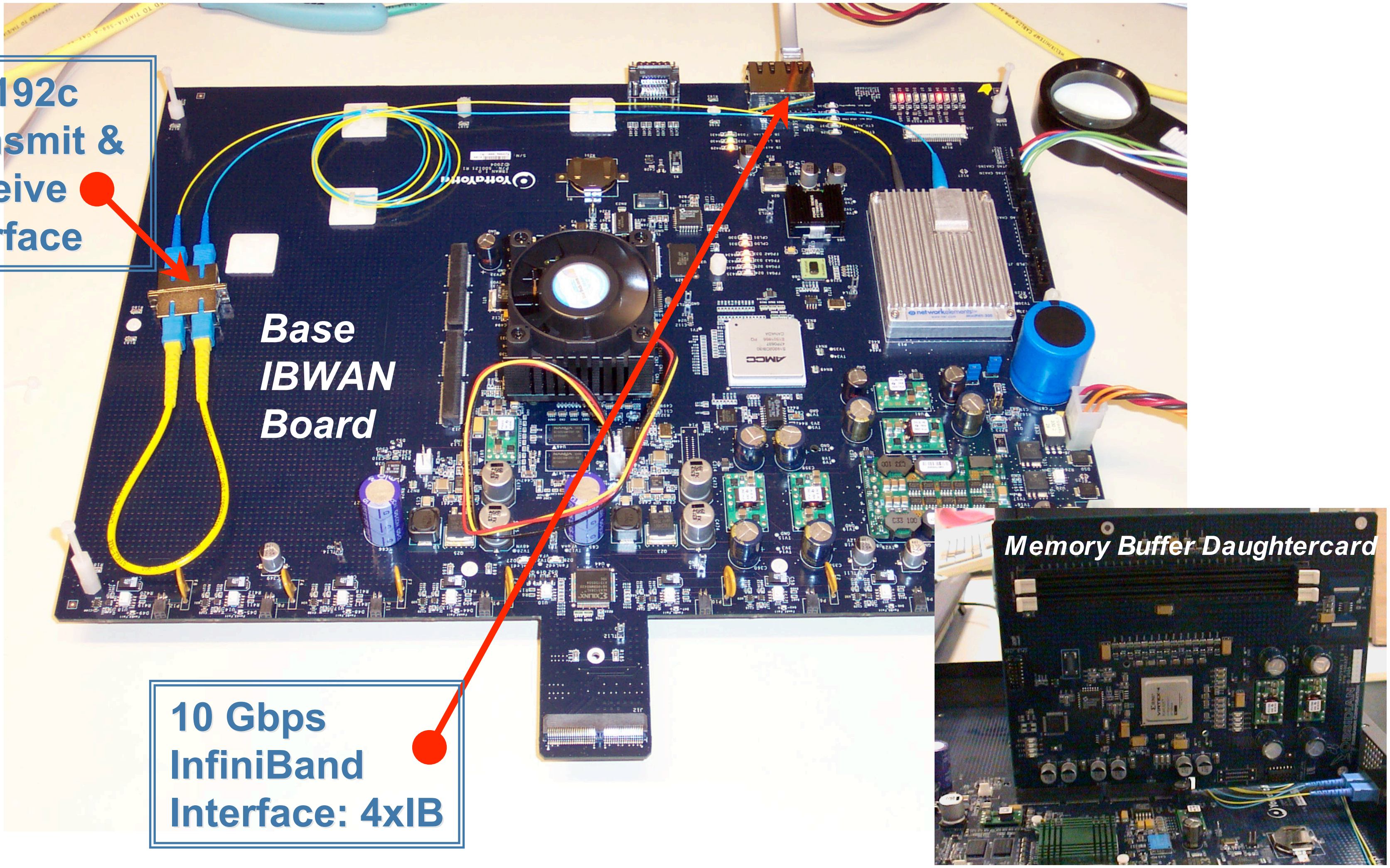
C. All protocol processing, memory access controls, etc. implemented in RNIC enabling complete off-load from the system.

# RDMA Infrastructure:
## *Solution Components*

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



# *IBWAN: Functional Prototype …*



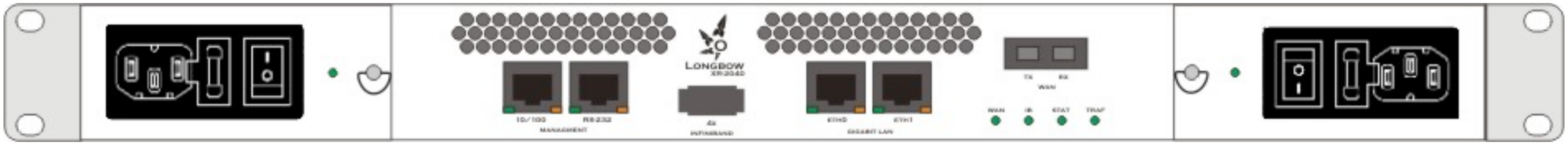
**OC-192c Transmit & Receive Interface**

**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
- 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 is 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.*

# *Toward Terabit Internetworking Functionality*

**4x IB WAN** *. . . Now*

*Point-to-point:*

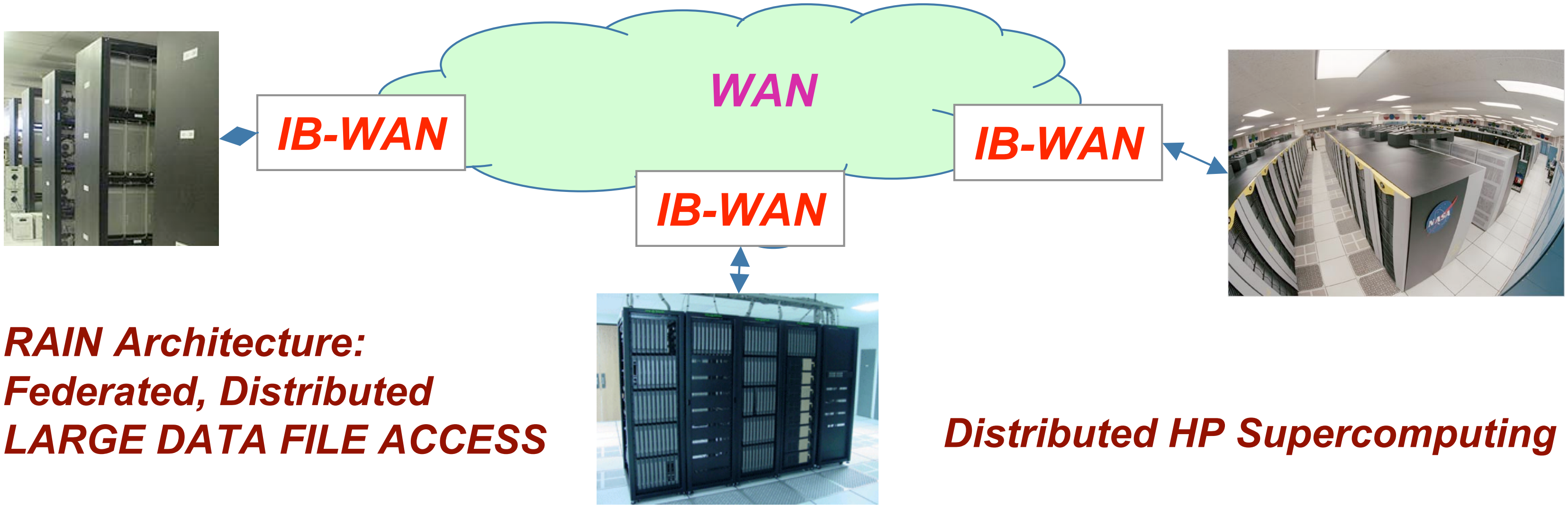
- *ATM/SONET (OC-192c)*
- *IPv6 POS (OC-192c)*

*Targeted: Nov 2005 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)*
- *JIT (dynamic)*



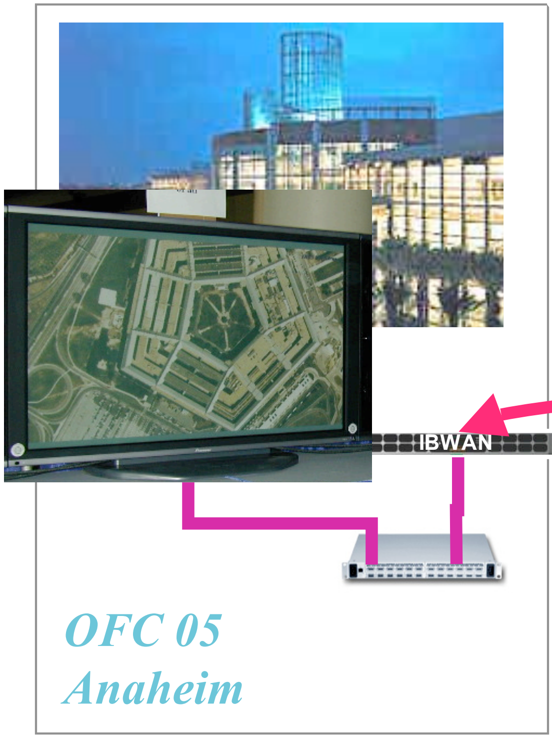
***RAIN Architecture:***
***Federated, Distributed***
***LARGE DATA FILE ACCESS***

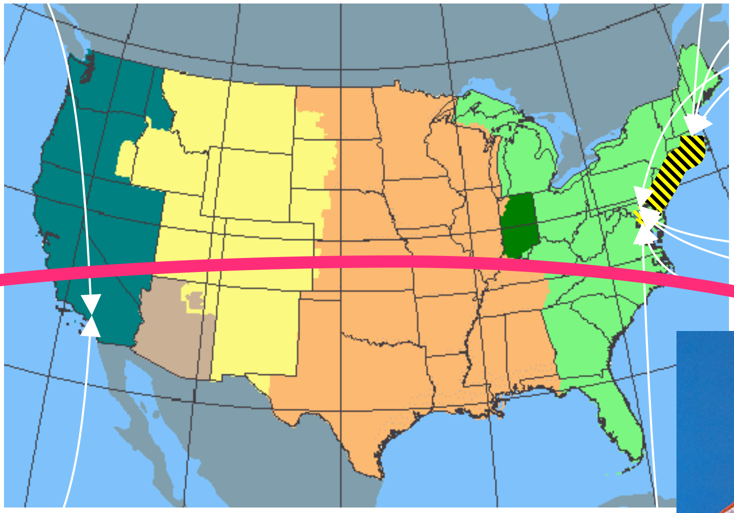
***Distributed HP Supercomputing***

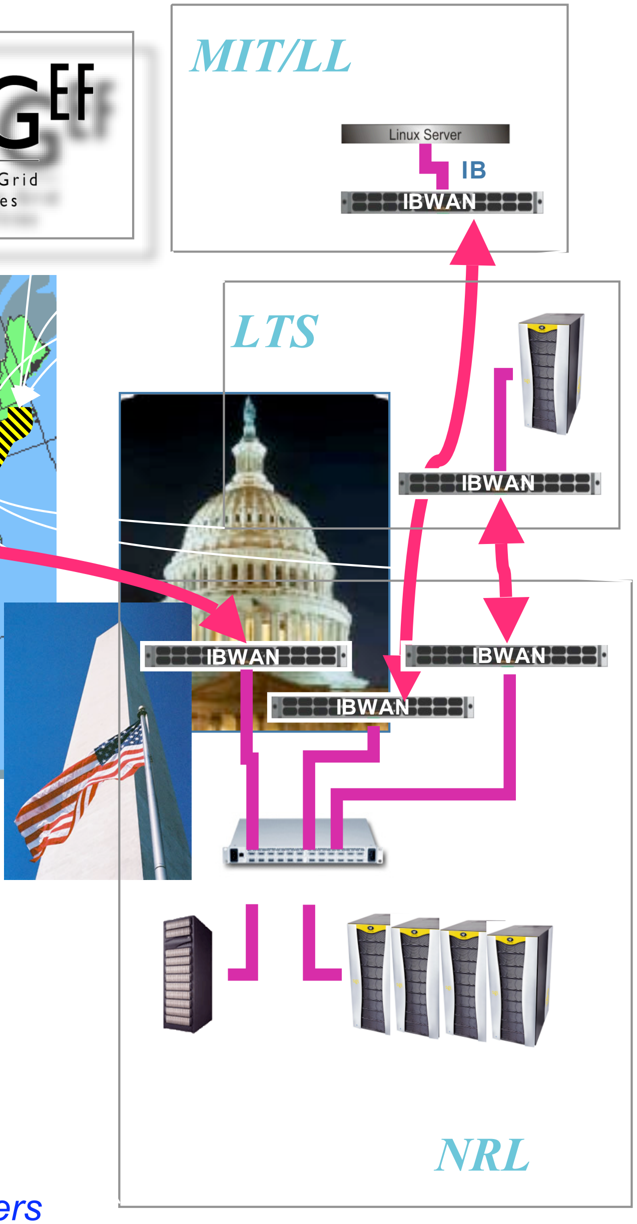
# InfiniBand Wide Area Networking at OFC/NFOEC 2005 …

## 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

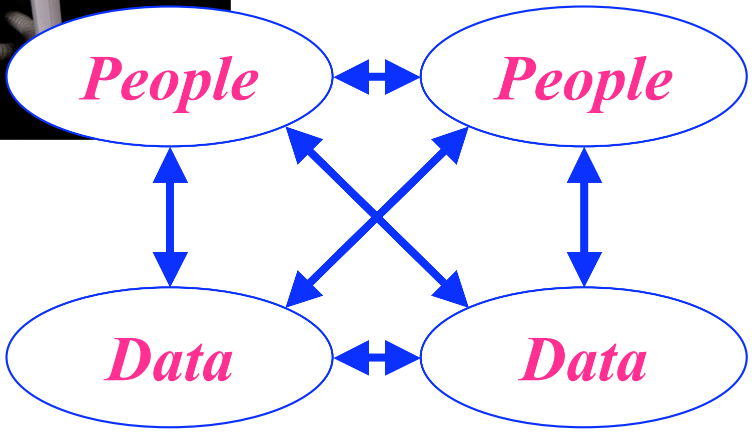
# Network Scaling Agenda . . .

| | 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<br>Provisioned | DYNAMIC<br>(GMPLS) | BURST/JIT<br>Just-in-time | |
| Control Plane | STATIC<br>Tunnel | DYNAMIC<br>SIP | SIP<br>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, . . . |

# ViPr: Video Presence Flexible Audio/Video Teleconferencing

… IPV6 based, SIP Control Plane, 14+1 Participants, "White Board" enabled, clear progressive HD video, echo-cancelled audio, touch controlled



A Desktop Collaboration Appliance:

**TRUE BROADBAND MULTI-PERSON**

## "DATA CONFERENCING"

*"… see the other guy sweat … realtime visual/audio/data collaboration, etc. "*

DEPARTMENT OF THE NAVY ★ NAVAL RESEARCH LABORATORY ★

*"Let's Roll!"*

*Thank You*

Center for Computational Science

*of the Naval Research Laboratory*