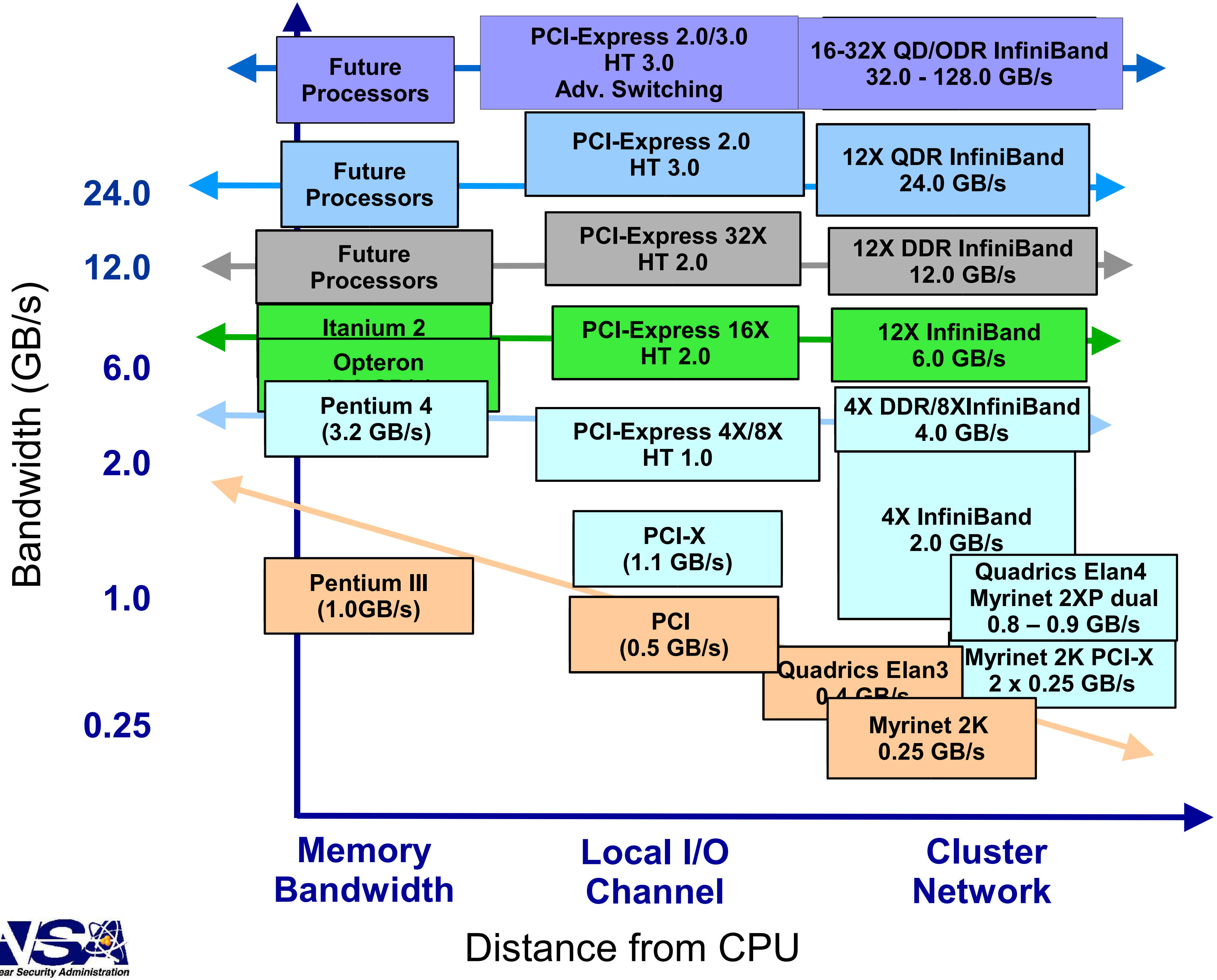
InfiniBand OpenIB Software PathForward Update

Matt Leininger, Curt Janssen, Mitch Sukalski (SNL)
Steve Poole, Ron Minnich, Mike Boorman, Rich Graham (LANL)
Bill Boas, Mark Seager, Terry Jones (LLNL)
Troy Benjegderes (Ames)

24 February 2005

InfiniBand Roadmap Tracks Future Processor and I/O Performance



Future Processors

PCI-Express 2.0/3.0
HT 3.0
Adv. Switching

16-32X QD/ODR InfiniBand
32.0 - 128.0 GB/s

Future Processors

PCI-Express 2.0
HT 3.0

12X QDR InfiniBand
24.0 GB/s

Future Processors

PCI-Express 32X
HT 2.0

12X DDR InfiniBand
12.0 GB/s

Itanium 2

Opteron

PCI-Express 16X
HT 2.0

12X InfiniBand
6.0 GB/s

Pentium 4
(3.2 GB/s)

PCI-Express 4X/8X
HT 1.0

4X DDR/8XInfiniBand
4.0 GB/s

4X InfiniBand
2.0 GB/s

PCI-X
(1.1 GB/s)

Quadrics Elan4
Myrinet 2XP dual
0.8 – 0.9 GB/s

Pentium III
(1.0GB/s)

PCI
(0.5 GB/s)

Quadrics Elan3
0.4 GB/s

Myrinet 2K PCI-X
2 x 0.25 GB/s

Myrinet 2K
0.25 GB/s

Goals of InfiniBand Software PathForward

- To accelerate the development of an InfiniBand software stack for HPC
  - High performance (high bandwidth, low latency)
  - Scalability
  - Robustness and Portability
  - Reliability
  - Manageability
  - Single open source SW stack supported across multiple system vendors
  - Integrate IB SW stack into mainline Linux kernel at kernel.org
  - Supported by Linux distributions (RedHat, SuSE, etc.)

How do we unite industry, open source, and HPC community behind these goals?

Form OpenIB Alliance to Unite Industry and Open Source Community

- Tri-labs are founding members of the OpenIB Alliance (www.openib.org)
- Other members include
  - Intel, Topspin, Voltaire, Mellanox, InfiniCon, Engenio, SGI, Linux Networx, NetApp, Oracle, Sun, Dell, Data Direct, and Veritas
- OpenIB is focusing on the goals in the previous slide
- Provide production ready Linux software solutions for HPC, data center, and scalable I/O.
- InfiniBand PathForward funds part of the work on the OpenIB stack that is focused on HPC – Voltaire, Topspin, and Intel
- Code is dual-licensed GPL/BSD

www.openib.org

OPEN IB ALLIANCE

National Nuclear Security Administration

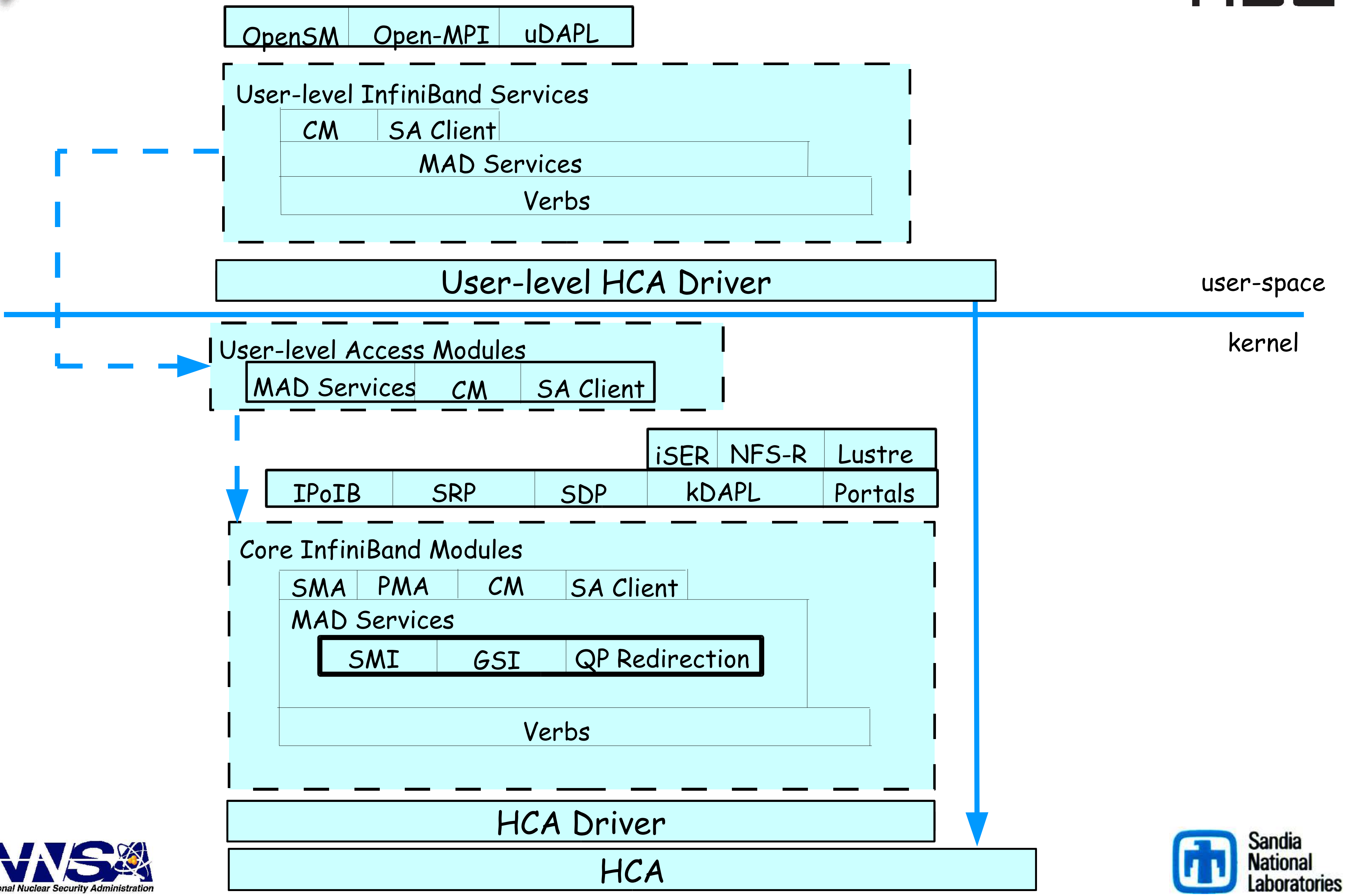
Sandia National Laboratories

Milestones of OpenIB PathForward

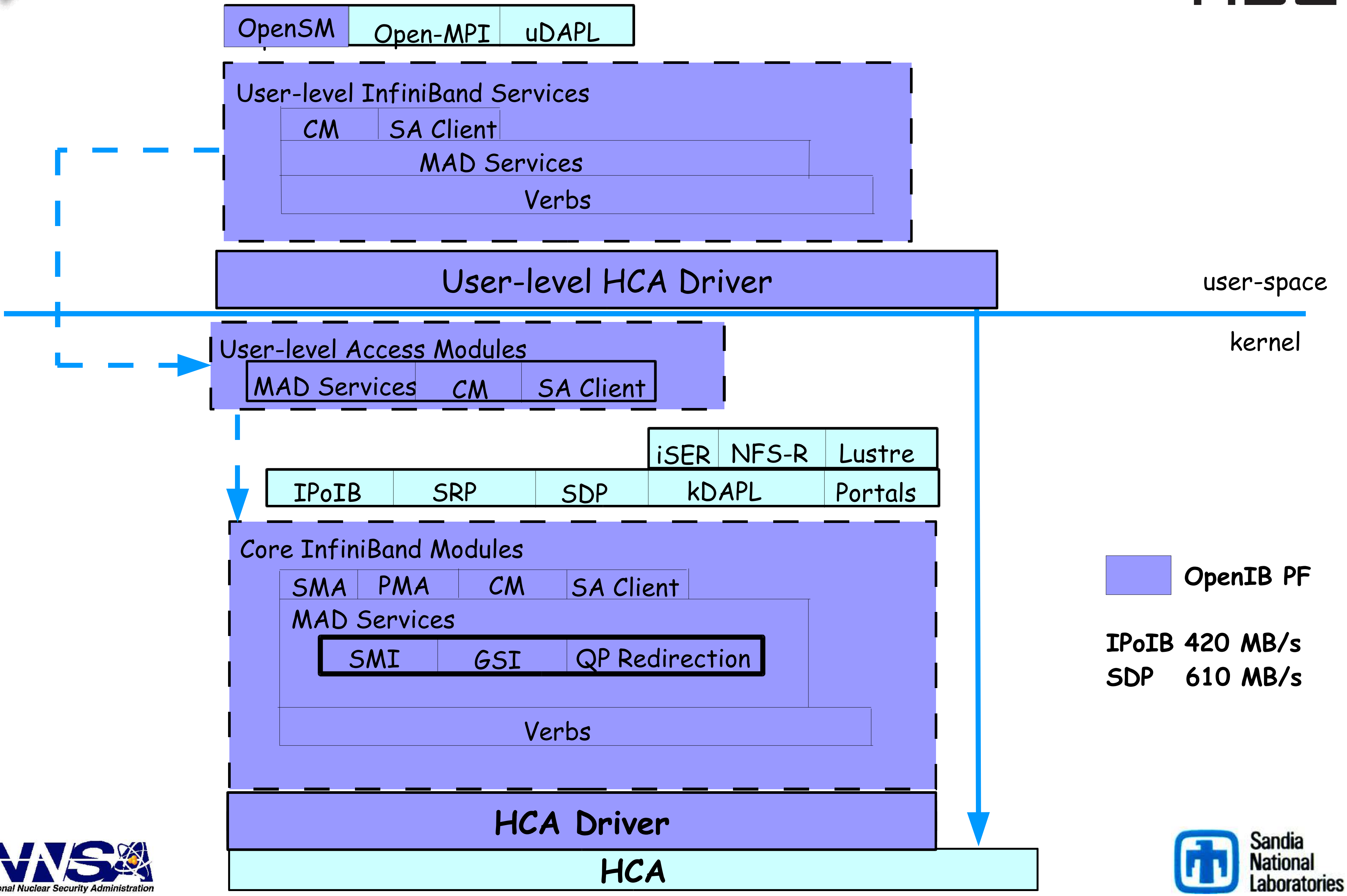
- FY05 Milestones completed
  - Project Plan & Project Design Review
  - OpenIB mthca driver and IPoIB accepted into Linux 2.6 kernel (2.6.11)
- FY05 Milestones (implementation)
  - Diagnostics
  - User-space verbs implemention for MPI
  - First release of OpenIB stack for HPC
- FY06 Milestones (optimization and some implementation)
  - Performance optimization of IB driver and user-space verbs
  - Complete IB host and network diagnostics
  - Continue to push OpenIB HPC stack updates to kernel.org
  - Optimized HPC routing
  - Network topology awareness
  - Official releases of OpenIB stack for HPC

Sandia and LANL working on OpenIB support in Open-MPI

OpenIB Stack Architecture



OpenIB Stack Architecture



OpenSM | Open-MPI | uDAPL

User-level InfiniBand Services

CM | SA Client

MAD Services

Verbs

User-level HCA Driver

user-space

kernel

User-level Access Modules

MAD Services | CM | SA Client

iSER | NFS-R | Lustre

IPoIB | SRP | SDP | kDAPL | Portals

Core InfiniBand Modules

SMA | PMA | CM | SA Client

MAD Services

SMI | GSI | QP Redirection

Verbs

HCA Driver

HCA

▨ OpenIB PF

IPoIB	420 MB/s
SDP	610 MB/s

Scope of OpenIB PathForward

- Funding OpenIB Alliance members Voltaire, Topspin, and Intel
- Focus on software components critical to HPC
- Access layer, diagnostics, subnet manager, scalability, performance, portability
- Achieved milestone with acceptance of the OpenIB driver into 2.6 kernel
- RedHat, SuSE, and other Linux distributions will start supporting IB
- InfiniBand support in Linux kernel is driving the wider adoption of InfiniBand
- Industry is leveraging the PF funding to drive IB adoption and to grow the OpenIB development community
- InfiniBand Device Driver – implementation
  - Use read/write instead of ioctl to avoid "big kernel" lock
  - Fast path ops require only a function call from (through function ptr) application to hardware access function
  - No context switches required in fast path
  - Interrupt-driven ops require kernel to wake up process; performance is limited by kernel interrupt service and scheduler latency
  - Code designed from start to reduce expensive ops (PCI reads, locking, cache misses)

OpenIB Developers Workshop Sets 2005 Agenda

- Held in Sonoma, CA Feb. 6-9, 2005
- Follow on workshop to DoE IB workshops over the last two years
- Organized by OpenIB Alliance (includes Tri-labs)
- Attendence has doubled each year from 2003 to 2005
  - 120 attendees from national labs, supercomputing centers, tier 1, 2, and 3, system providers, InfiniBand and storage industries, embedded systems, academia, Linux distributions, and database software.
- Over 30 developers covering every component of the OpenIB stack
- Kernel – HCA driver, Verbs, MAD services and core modules, IPoIB, SRP, SDP, kDAPL, iSER, Portals, NFS-RDMA, and user level access modules
- User-space – user level HCA driver, MAD services, connection manager, subnet manager (OpenSM), MPI (Open-MPI, MVAPICH), and uDAPL
- OpenIB SW stack development is expanding beyond the PF efforts
- InfiniBand is now seen as a player in a wider market
  - Oracle and IBM are requiring OpenIB stack in distros for DB applications
  - RedHat plans to have OpenIB code in RH Enterprise Linux 4 updates

OpenIB PathForward Summary

- PathForward is accelerating the development of an open source InfiniBand software stack with HPC capabilities that is supported by industry, Linux distributions, and the open source communities
- OpenIB Alliance aligns these communities behind a single IB stack
- Open source & open development required to get into Linux kernel
- Achieved first major milestone with OpenIB driver and stack being accepted into 2.6 Linux kernel at kernel.org
- Code flows from OpenIB to kernel.org to Linux distributions
- Sandia and LANL are implementing OpenIB support in Open-MPI
- OpenIB Developers Workshops every 6 months
- Code is available for download at www.openib.org. Join OpenIB mail list to see progress
- PathForward funding is ensuring that ASC's requirements stay at the forefront of OpenIB development

For more information

www.openib.org

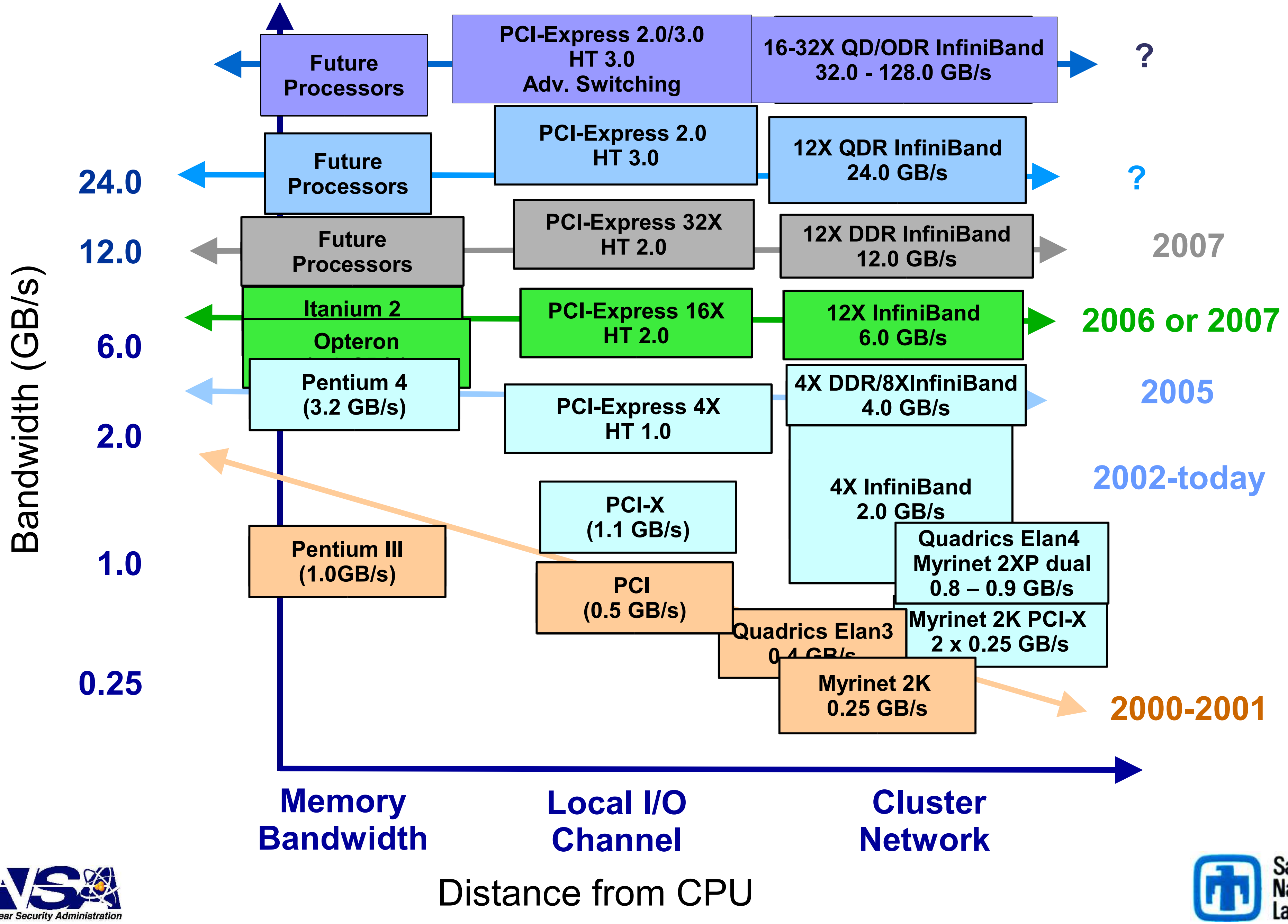
mlleini@sandia.gov

Backup Slides

InfiniBand Motivation

- Leverage commodity InfiniBand (IB) to benefit from the economies of scale
  - Much wider business opportunities than NNSA (or even HPC)
  - Cost for 4X/8X IB $1010/port dropping to $520-670/port (Myrinet $1100/port, Quadrics $2950/port)
- IB has the potential to meet or exceed the performance of proprietary interconnects
  - IB has roadmap from 20 Gbps (2 GB/s) to 240 Gbps (24 GB/s), and 1-4 us latency
- Make IB viable for next platform procurements
- IB software will not meet HPC performance and timeline requirements w/o PF
- Require OS software stack for HPC to reap the full benefits of IB
- IB is the top priority of the OSSODA community
- DOE IB Workshop concluded that we focus on OS HPC IB development
  1. ***Multi-vendor*** **IB Software Stack for HPC integrated into the Linux kernel**
  2. **Scalable InfiniBand Diagnostic and Management Tools**
  3. **Scalable System Software and MPI Middleware**
  4. **Platform Independence and Portability**
  5. **Latency Reduction**
- Use OSSODA to accelerate these IB HPC requirements and Form OpenIB Alliance

InfiniBand Roadmap Tracks Future Processor and I/O Performance



Bandwidth (GB/s)

24.0

12.0

6.0

2.0

1.0

0.25

Future Processors

PCI-Express 2.0/3.0
HT 3.0
Adv. Switching

16-32X QD/ODR InfiniBand
32.0 - 128.0 GB/s

?

Future Processors

PCI-Express 2.0
HT 3.0

12X QDR InfiniBand
24.0 GB/s

?

Future Processors

PCI-Express 32X
HT 2.0

12X DDR InfiniBand
12.0 GB/s

2007

Itanium 2

Opteron

PCI-Express 16X
HT 2.0

12X InfiniBand
6.0 GB/s

2006 or 2007

Pentium 4
(3.2 GB/s)

PCI-Express 4X
HT 1.0

4X DDR/8XInfiniBand
4.0 GB/s

2005

PCI-X
(1.1 GB/s)

4X InfiniBand
2.0 GB/s

2002-today

Quadrics Elan4
Myrinet 2XP dual
0.8 – 0.9 GB/s

Pentium III
(1.0GB/s)

PCI
(0.5 GB/s)

Quadrics Elan3
0.4 GB/s

Myrinet 2K PCI-X
2 x 0.25 GB/s

Myrinet 2K
0.25 GB/s

2000-2001

Memory Bandwidth

Local I/O Channel

Cluster Network

Distance from CPU