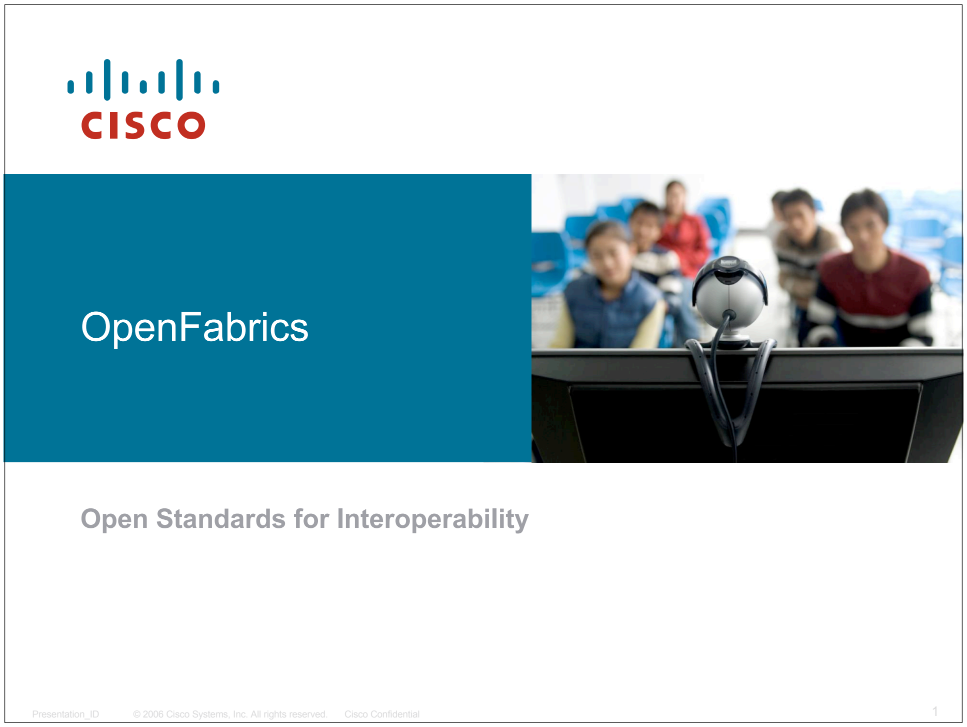
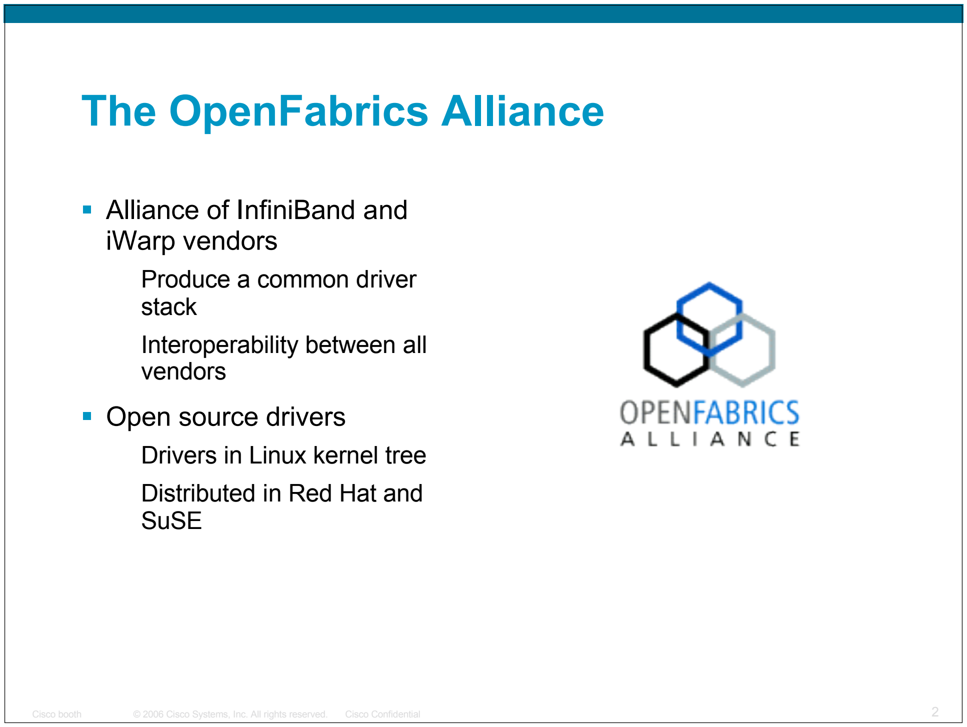


CISCO

OpenFabrics

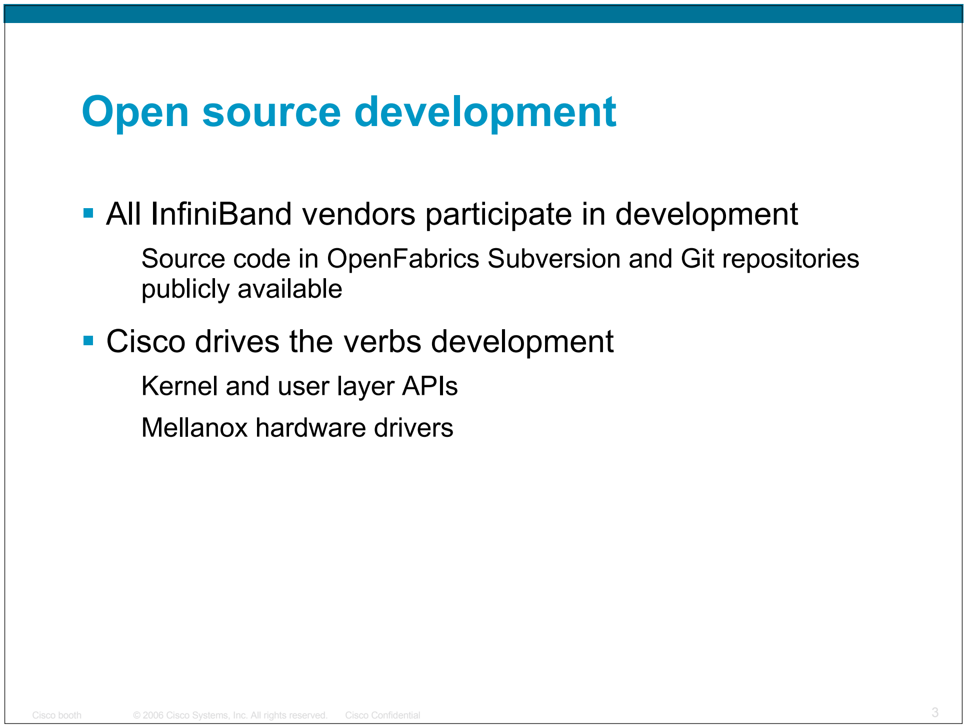
Open Standards for Interoperability



The OpenFabrics Alliance

- Alliance of InfiniBand and iWarp vendors
  - Produce a common driver stack
  - Interoperability between all vendors
- Open source drivers
  - Drivers in Linux kernel tree
  - Distributed in Red Hat and SuSE

OPENFABRICS ALLIANCE



Open source development

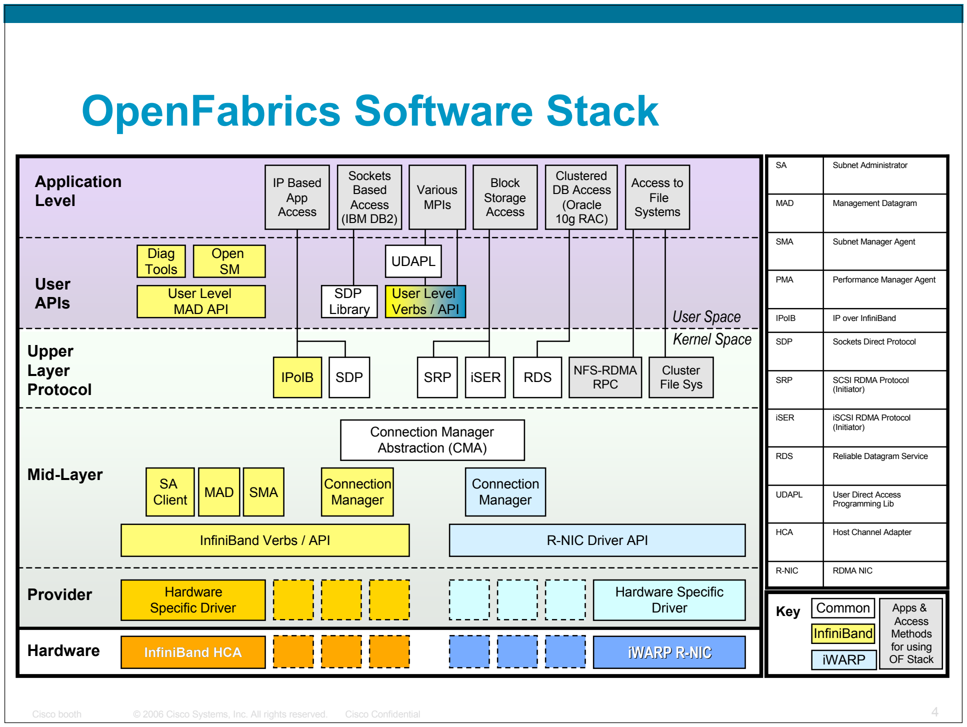
- All InfiniBand vendors participate in development

  Source code in OpenFabrics Subversion and Git repositories publicly available

- Cisco drives the verbs development

  Kernel and user layer APIs

  Mellanox hardware drivers



OpenFabrics Software Stack

Application Level

IP Based App Access

Sockets Based Access (IBM DB2)

Various MPIs

Block Storage Access

Clustered DB Access (Oracle 10g RAC)

Access to File Systems

User APIs

Diag Tools

Open SM

User Level MAD API

UDAPL

SDP Library

User Level Verbs / API

User Space

Kernel Space

Upper Layer Protocol

IPoIB

SDP

SRP

iSER

RDS

NFS-RDMA RPC

Cluster File Sys

Mid-Layer

Connection Manager Abstraction (CMA)

SA Client

MAD

SMA

Connection Manager

Connection Manager

InfiniBand Verbs / API

R-NIC Driver API

Provider

Hardware Specific Driver

Hardware Specific Driver

Hardware

InfiniBand HCA

iWARP R-NIC

SA	Subnet Administrator
MAD	Management Datagram
SMA	Subnet Manager Agent
PMA	Performance Manager Agent
IPoIB	IP over InfiniBand
SDP	Sockets Direct Protocol
SRP	SCSI RDMA Protocol (Initiator)
iSER	iSCSI RDMA Protocol (Initiator)
RDS	Reliable Datagram Service
UDAPL	User Direct Access Programming Lib
HCA	Host Channel Adapter
R-NIC	RDMA NIC

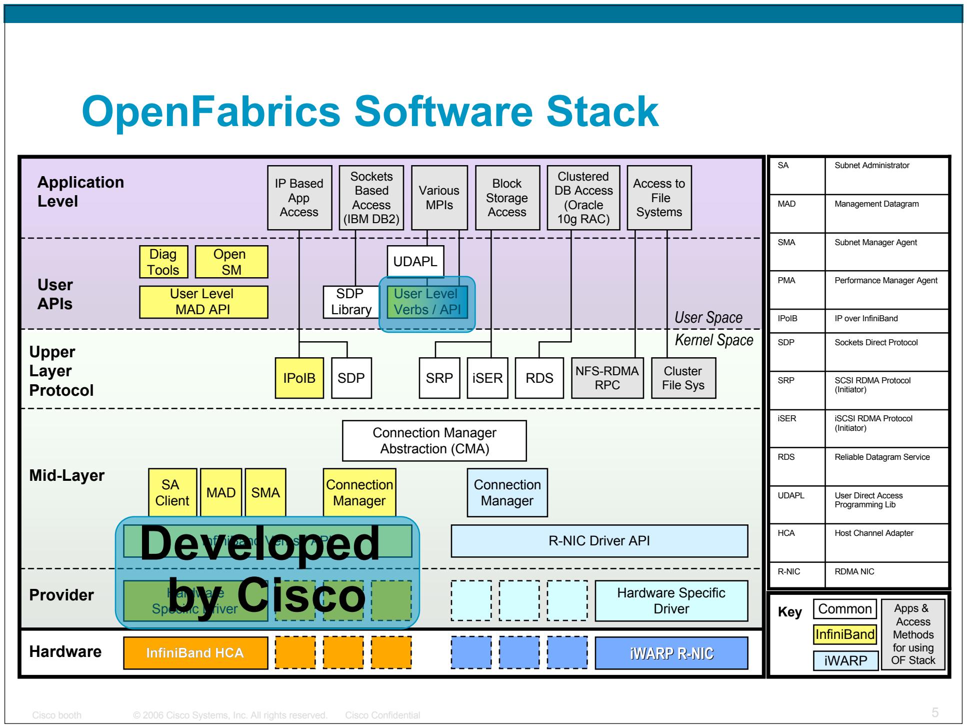
Key

Common

InfiniBand

iWARP

Apps & Access Methods for using OF Stack



OpenFabrics Software Stack

Application Level: IP Based App Access; Sockets Based Access (IBM DB2); Various MPIs; Block Storage Access; Clustered DB Access (Oracle 10g RAC); Access to File Systems

User APIs: Diag Tools; Open SM; User Level MAD API; SDP Library; UDAPL; User Level Verbs / API

User Space

Kernel Space

Upper Layer Protocol: IPoIB; SDP; SRP; iSER; RDS; NFS-RDMA RPC; Cluster File Sys

Mid-Layer: Connection Manager Abstraction (CMA); SA Client; MAD; SMA; Connection Manager; Connection Manager; InfiniBand Verbs / API; R-NIC Driver API

Provider: Hardware Specific Driver; Hardware Specific Driver

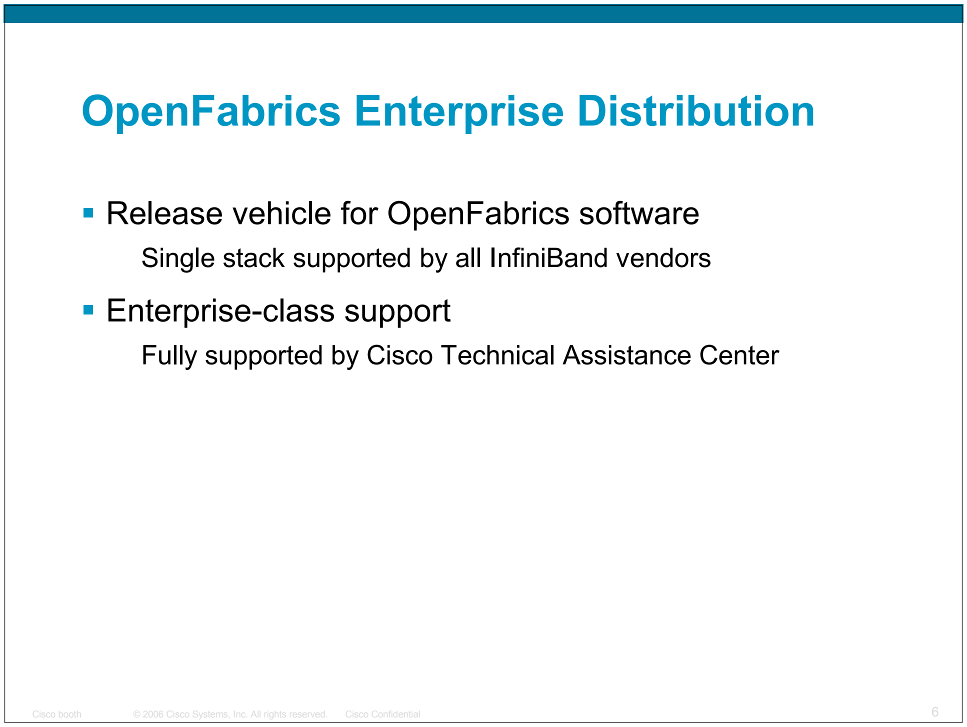
Hardware: InfiniBand HCA; iWARP R-NIC

Developed by Cisco

SA	Subnet Administrator
MAD	Management Datagram
SMA	Subnet Manager Agent
PMA	Performance Manager Agent
IPoIB	IP over InfiniBand
SDP	Sockets Direct Protocol
SRP	SCSI RDMA Protocol (Initiator)
iSER	iSCSI RDMA Protocol (Initiator)
RDS	Reliable Datagram Service
UDAPL	User Direct Access Programming Lib
HCA	Host Channel Adapter
R-NIC	RDMA NIC

Key: Common; InfiniBand; iWARP; Apps & Access Methods for using OF Stack

Cisco booth © 2006 Cisco Systems, Inc. All rights reserved. Cisco Confidential 5



OpenFabrics Enterprise Distribution

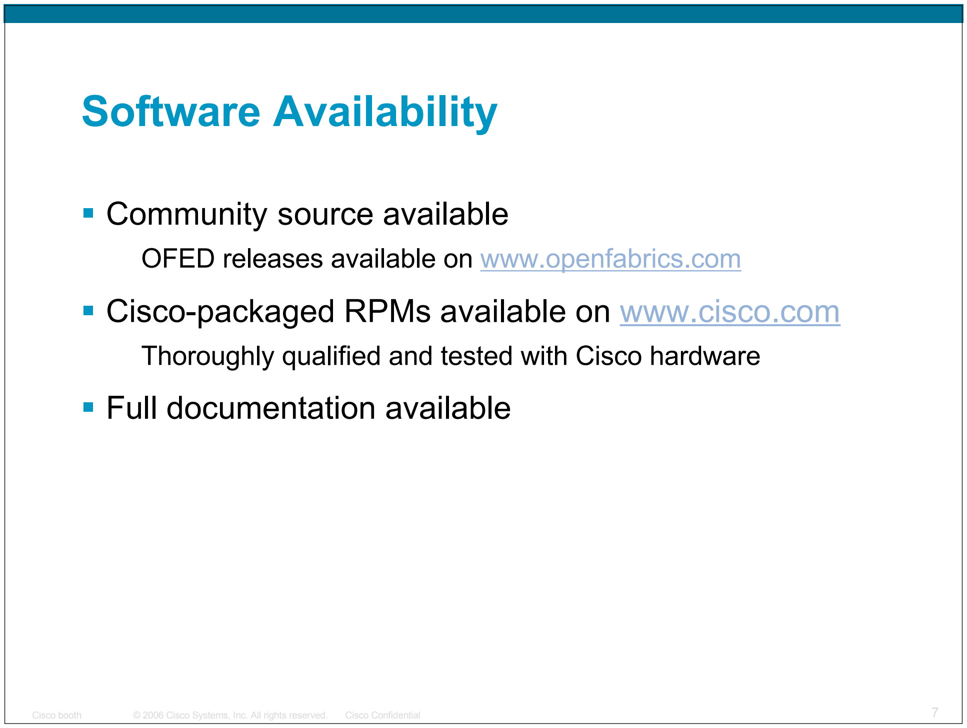
- Release vehicle for OpenFabrics software

  Single stack supported by all InfiniBand vendors

- Enterprise-class support

  Fully supported by Cisco Technical Assistance Center

Cisco booth © 2006 Cisco Systems, Inc. All rights reserved. Cisco Confidential 6



Software Availability

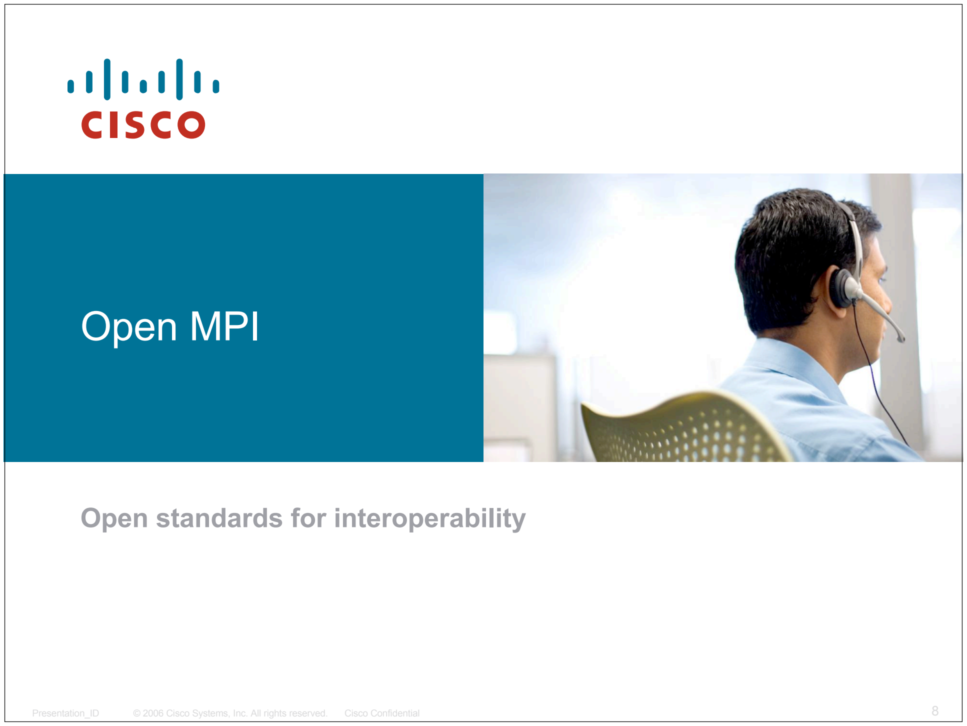
- Community source available

  OFED releases available on www.openfabrics.com

- Cisco-packaged RPMs available on www.cisco.com

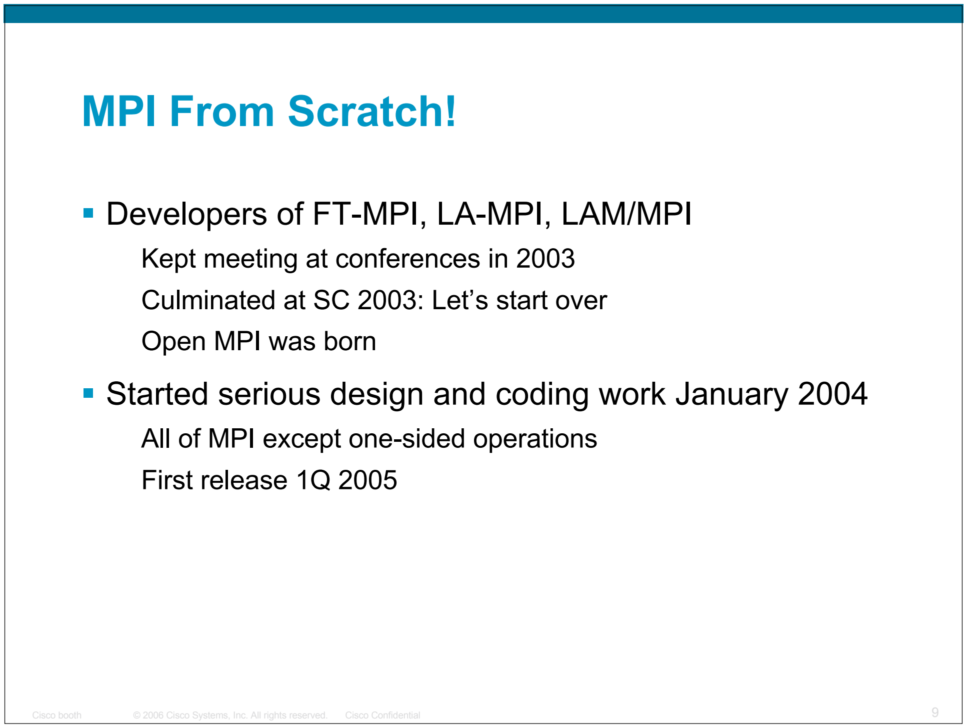
  Thoroughly qualified and tested with Cisco hardware

- Full documentation available



Open MPI

Open standards for interoperability



MPI From Scratch!

- Developers of FT-MPI, LA-MPI, LAM/MPI

  Kept meeting at conferences in 2003

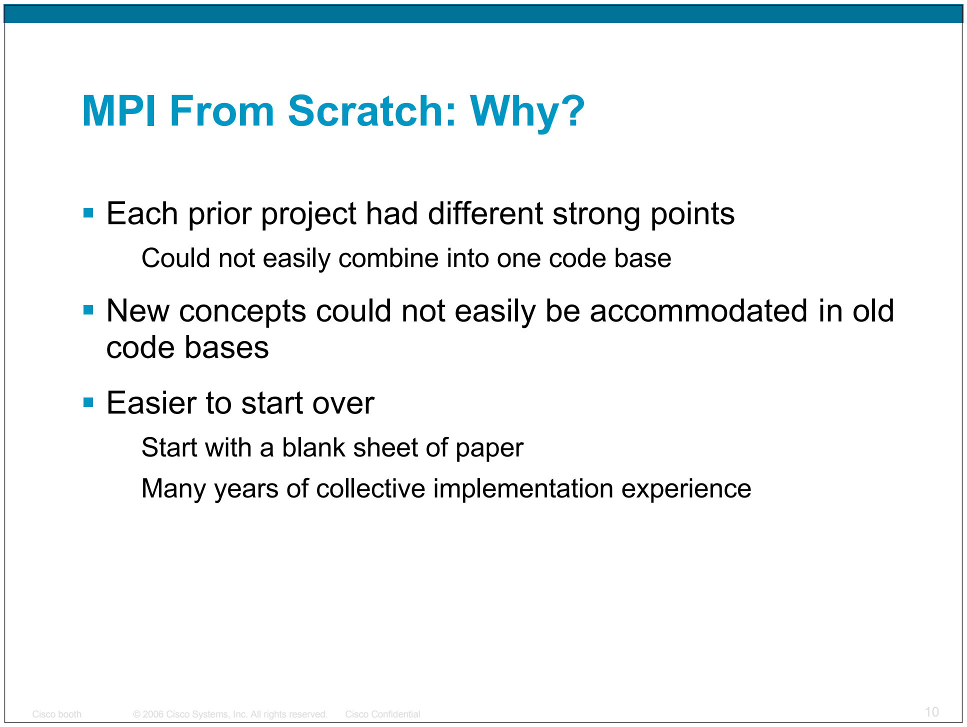
  Culminated at SC 2003: Let’s start over

  Open MPI was born

- Started serious design and coding work January 2004

  All of MPI except one-sided operations

  First release 1Q 2005



MPI From Scratch: Why?

- Each prior project had different strong points

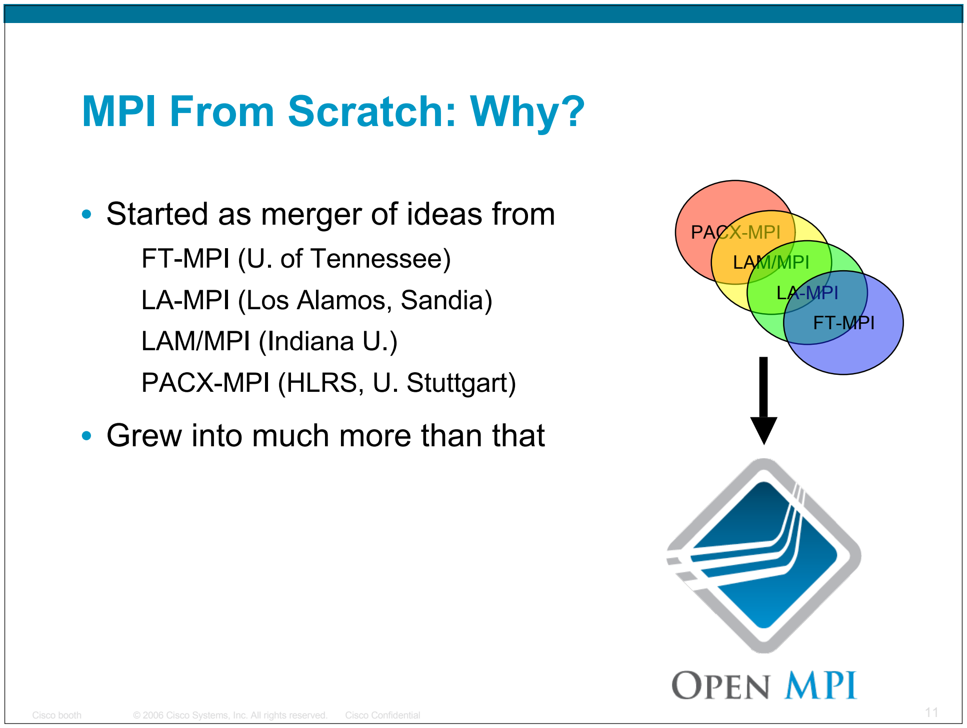
  Could not easily combine into one code base

- New concepts could not easily be accommodated in old code bases

- Easier to start over

  Start with a blank sheet of paper

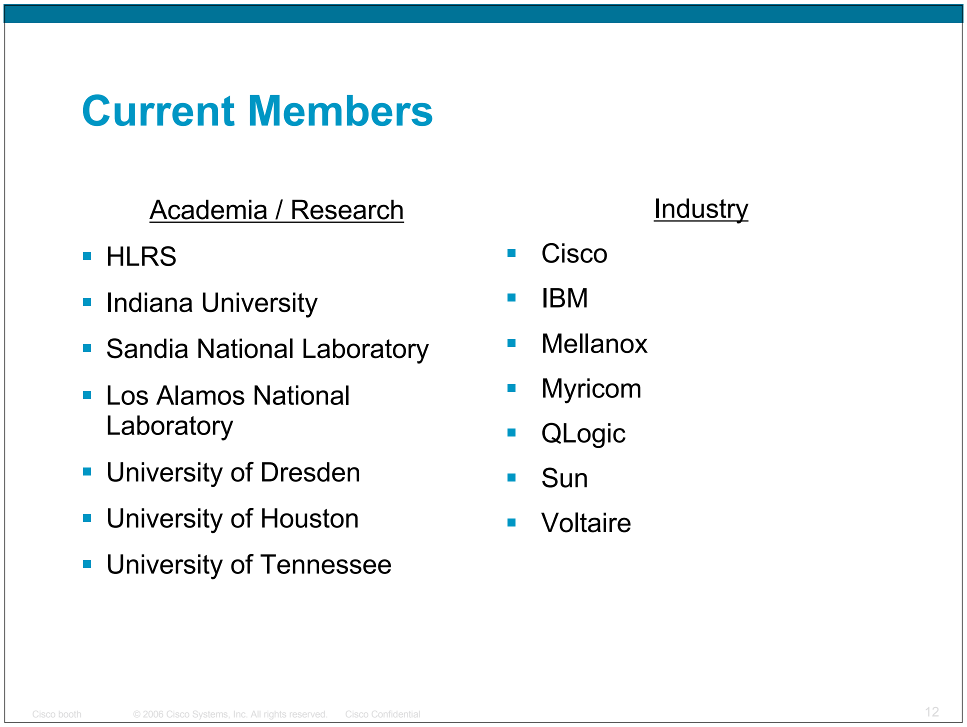
  Many years of collective implementation experience



MPI From Scratch: Why?

- Started as merger of ideas from
  - FT-MPI (U. of Tennessee)
  - LA-MPI (Los Alamos, Sandia)
  - LAM/MPI (Indiana U.)
  - PACX-MPI (HLRS, U. Stuttgart)
- Grew into much more than that

PACX-MPI
LAM/MPI
LA-MPI
FT-MPI
OPEN MPI



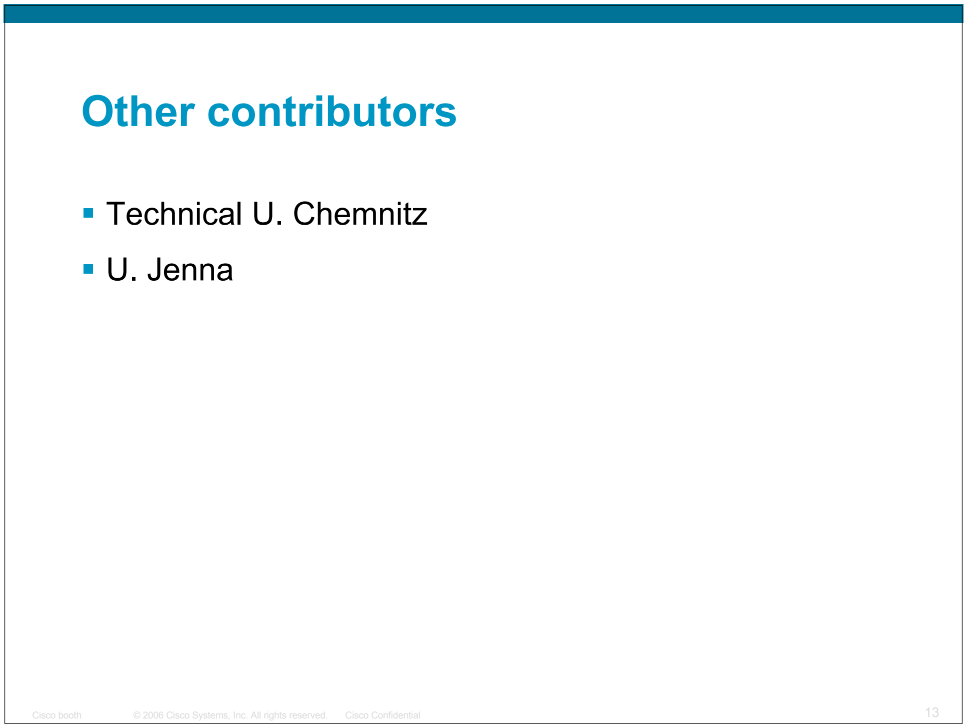
Current Members

Academia / Research

- HLRS
- Indiana University
- Sandia National Laboratory
- Los Alamos National Laboratory
- University of Dresden
- University of Houston
- University of Tennessee

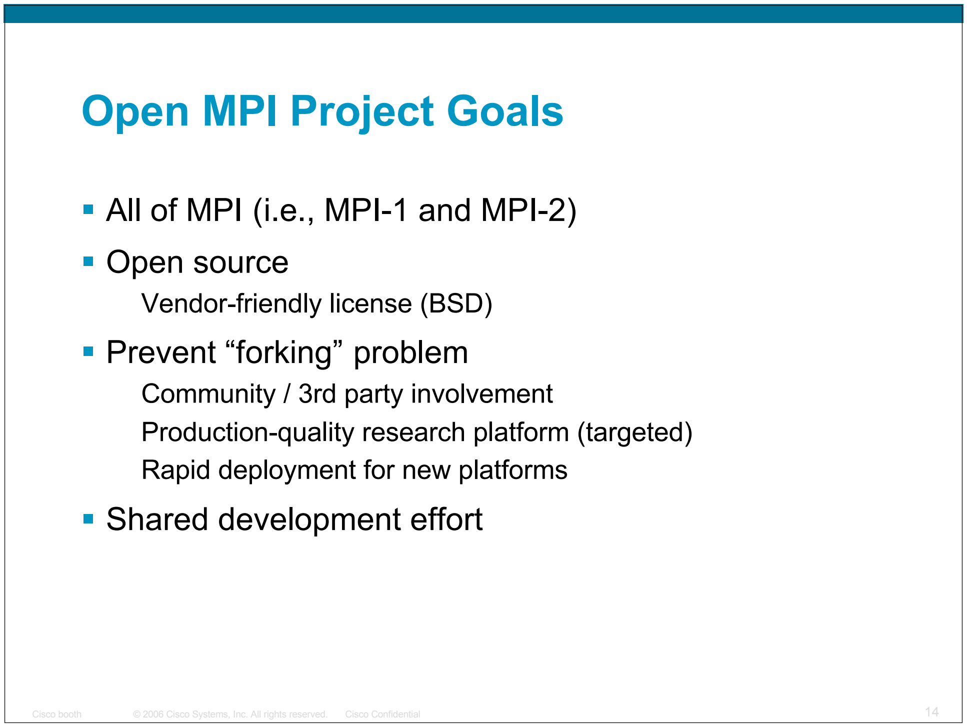
Industry

- Cisco
- IBM
- Mellanox
- Myricom
- QLogic
- Sun
- Voltaire



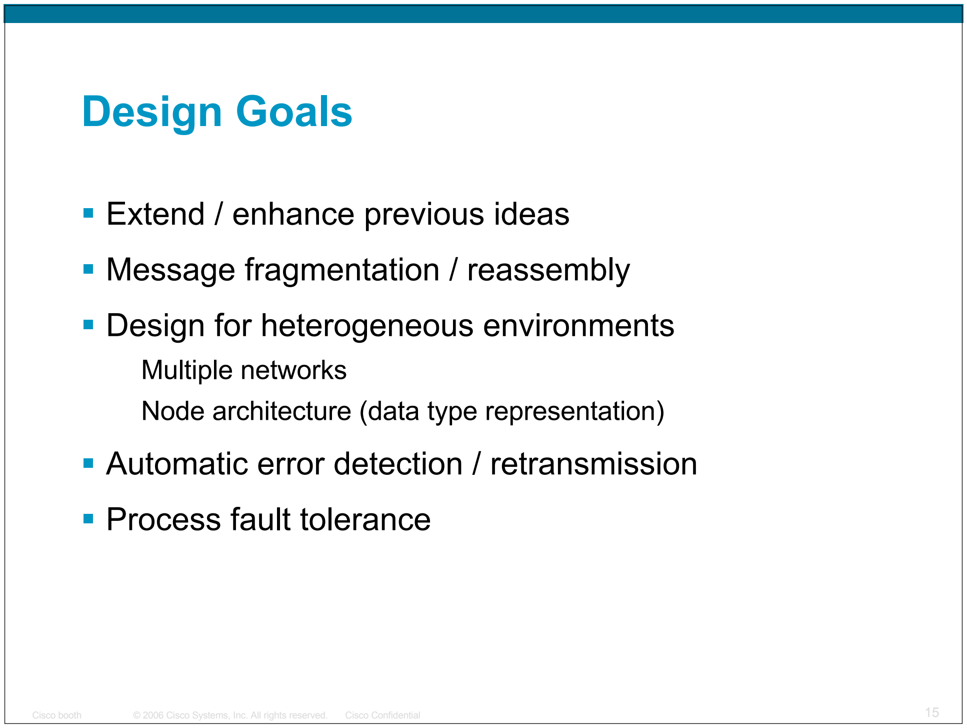
Other contributors

- Technical U. Chemnitz
- U. Jenna



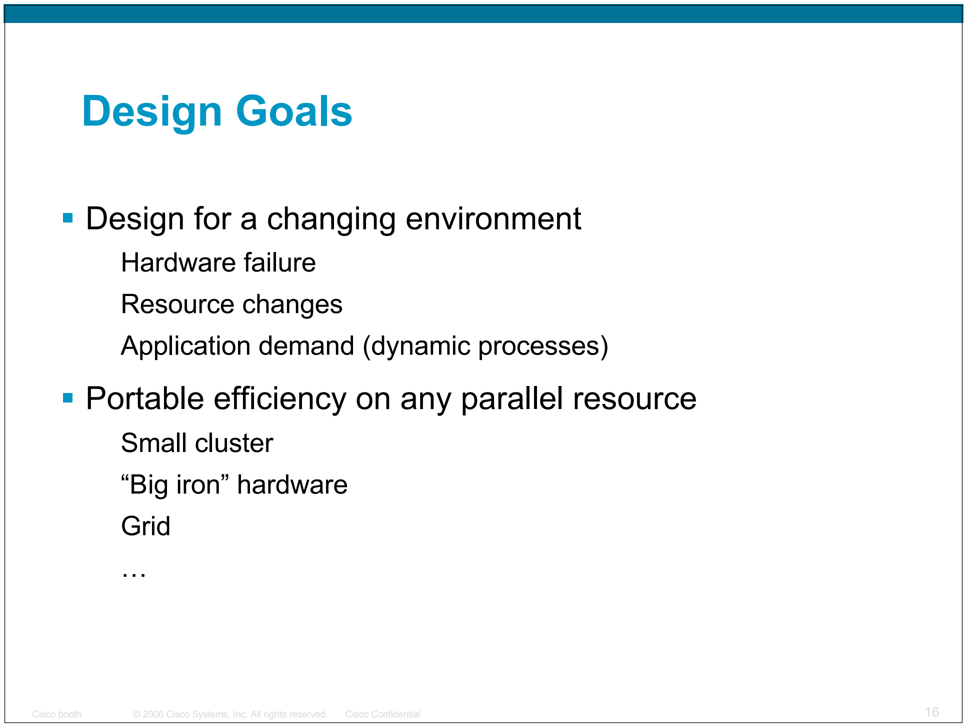
Open MPI Project Goals

- All of MPI (i.e., MPI-1 and MPI-2)
- Open source
  - Vendor-friendly license (BSD)
- Prevent “forking” problem
  - Community / 3rd party involvement
  - Production-quality research platform (targeted)
  - Rapid deployment for new platforms
- Shared development effort



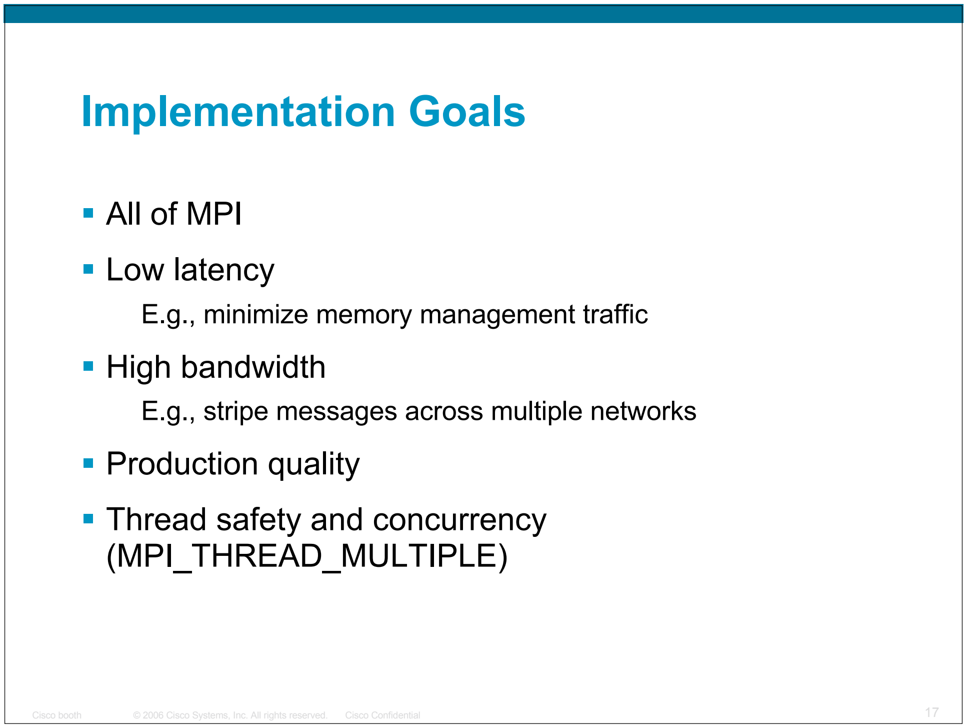
Design Goals

- Extend / enhance previous ideas
- Message fragmentation / reassembly
- Design for heterogeneous environments
  - Multiple networks
  - Node architecture (data type representation)
- Automatic error detection / retransmission
- Process fault tolerance



Design Goals

- Design for a changing environment
  - Hardware failure
  - Resource changes
  - Application demand (dynamic processes)
- Portable efficiency on any parallel resource
  - Small cluster
  - “Big iron” hardware
  - Grid
  - …

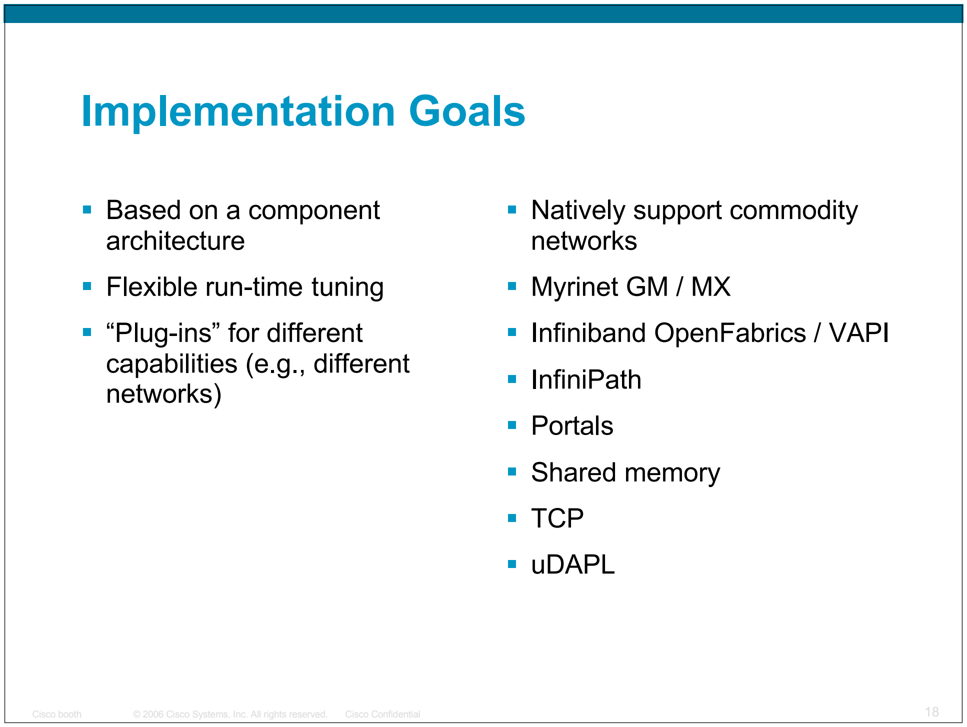


Implementation Goals

- All of MPI
- Low latency

  E.g., minimize memory management traffic
- High bandwidth

  E.g., stripe messages across multiple networks
- Production quality
- Thread safety and concurrency (MPI_THREAD_MULTIPLE)



Implementation Goals

- Based on a component architecture
- Flexible run-time tuning
- “Plug-ins” for different capabilities (e.g., different networks)

- Natively support commodity networks
- Myrinet GM / MX
- Infiniband OpenFabrics / VAPI
- InfiniPath
- Portals
- Shared memory
- TCP
- uDAPL

Current Status

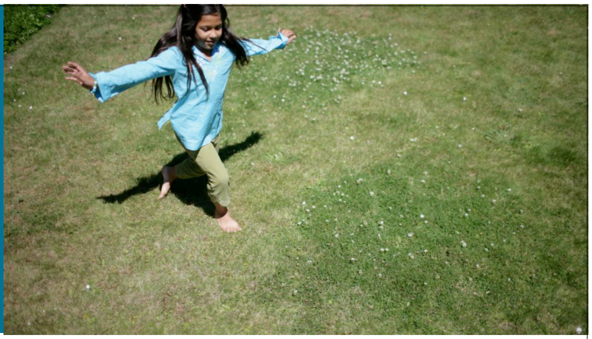
- Open MPI v1.1.2 current stable release

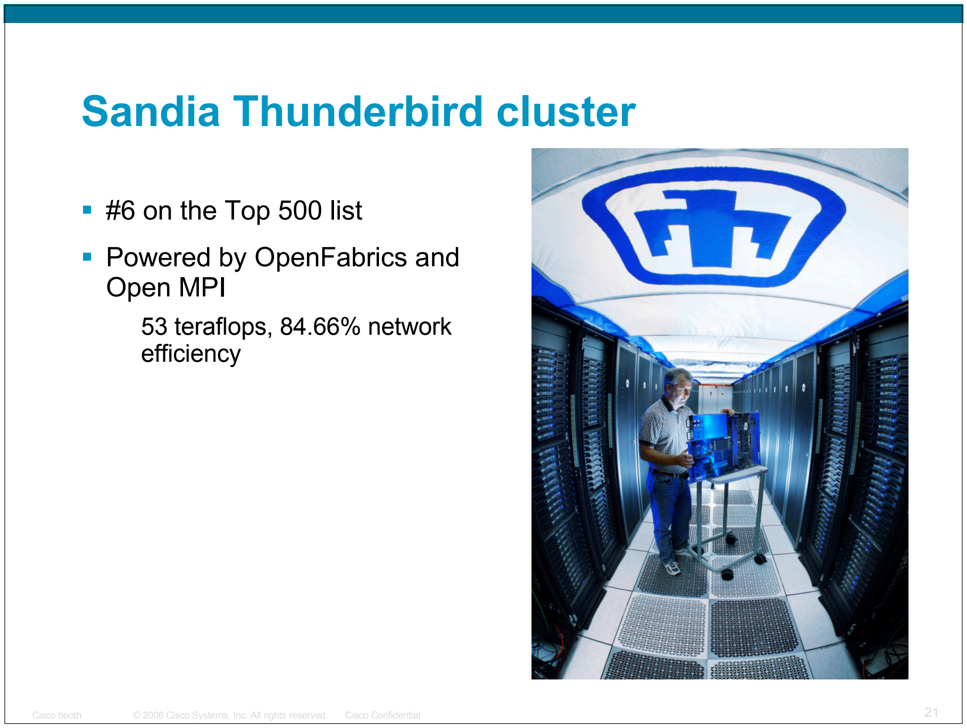
  Included in OFED distributions

- Open MPI v1.2b1 available for preview

  http://www.open-mpi.org/

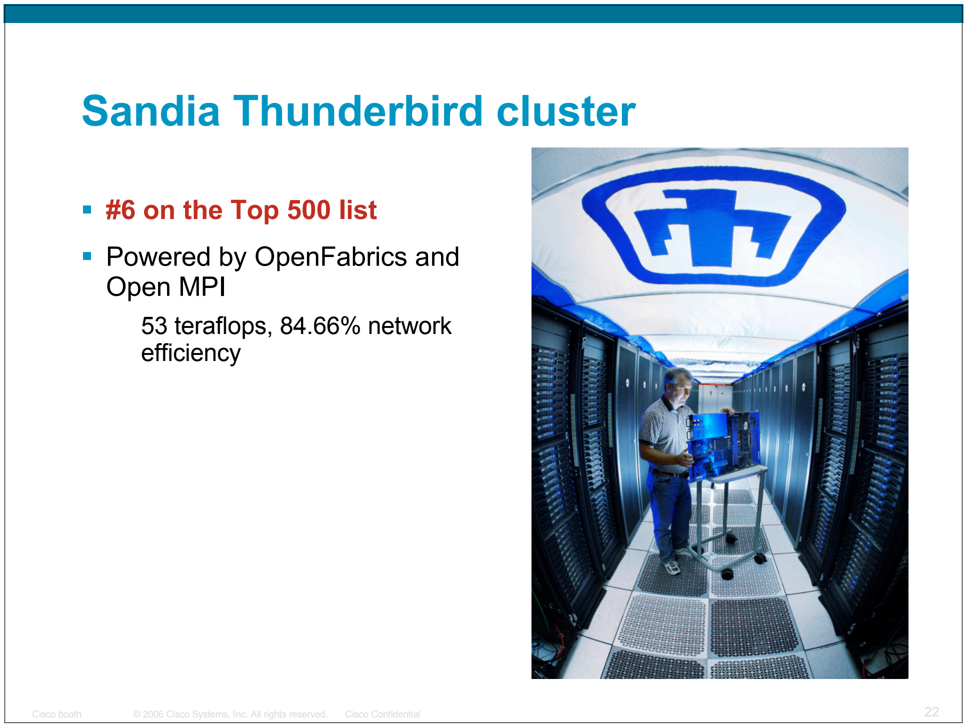
The Power of Open Standards





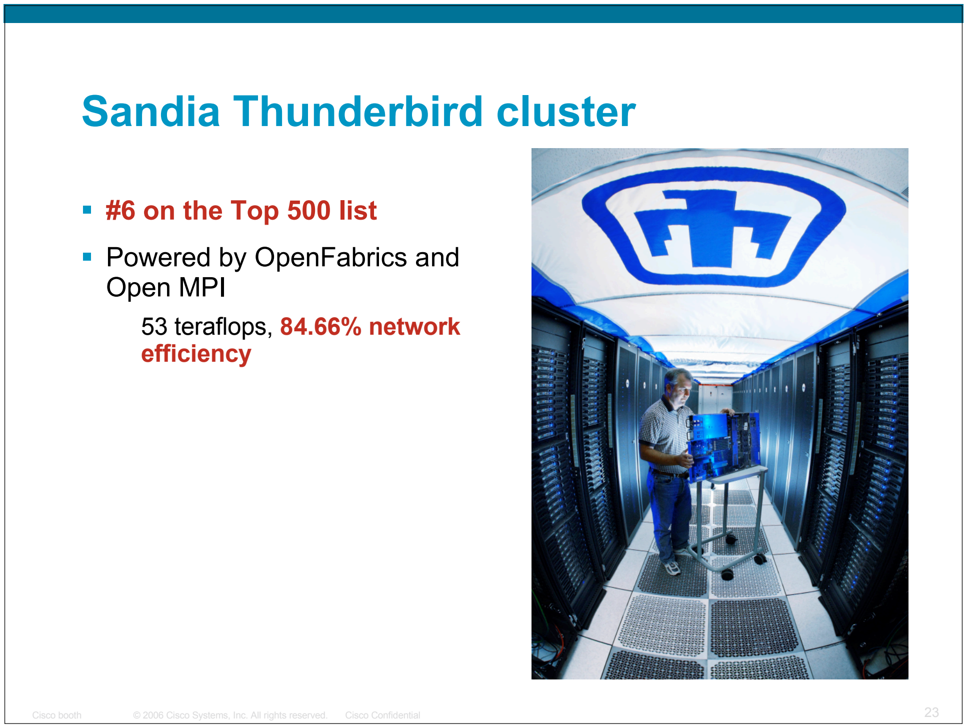
Sandia Thunderbird cluster

- #6 on the Top 500 list
- Powered by OpenFabrics and Open MPI
  - 53 teraflops, 84.66% network efficiency



Sandia Thunderbird cluster

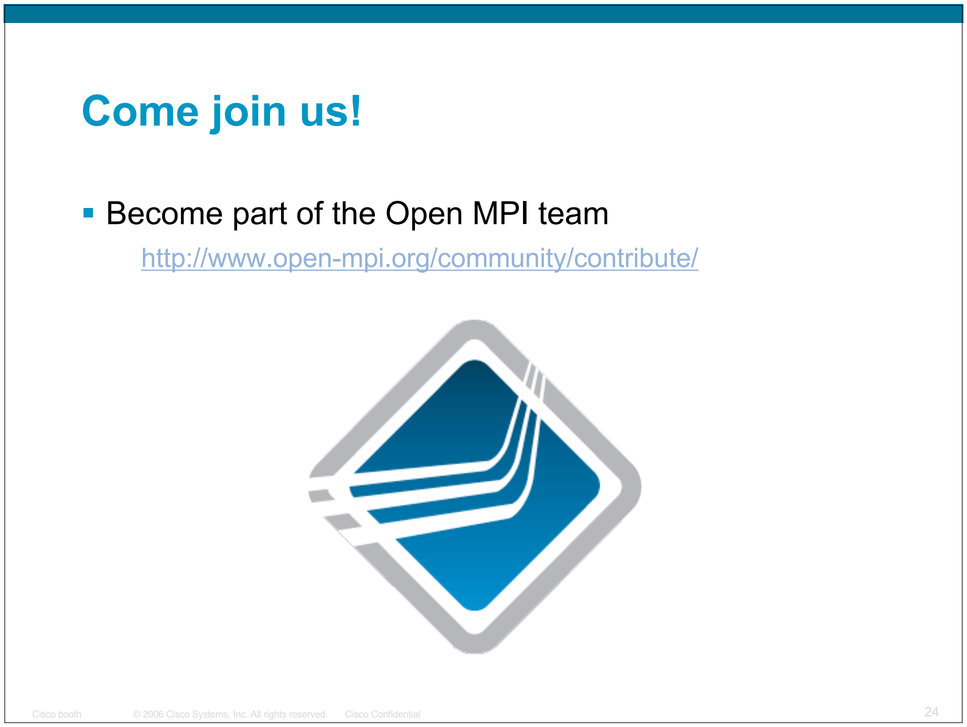
- **#6 on the Top 500 list**
- Powered by OpenFabrics and Open MPI
  - 53 teraflops, 84.66% network efficiency



Sandia Thunderbird cluster

- **#6 on the Top 500 list**
- Powered by OpenFabrics and Open MPI

  53 teraflops, **84.66% network efficiency**



Come join us!

- Become part of the Open MPI team

  http://www.open-mpi.org/community/contribute/