Cray Inc.

Adaptive Supercomputing

Matthew O'Keefe
Storage Architect

SOS-10 Data Intensive Computing Panel

7 March 2006

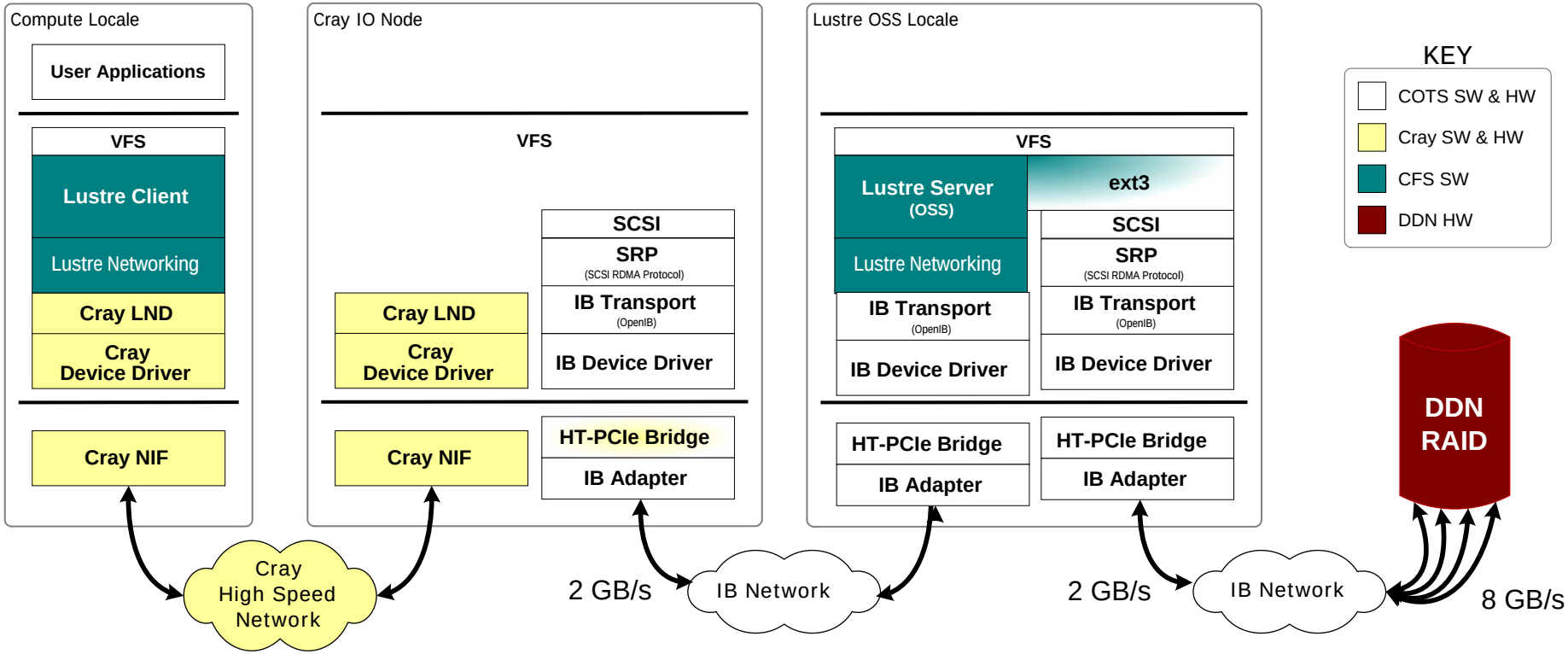
Cray Storage Directions

- Adaptive Supercomputing
  - Use the right storage to meet a particular application/data center requirement
- Speeds and Feeds
  - Terabyte-per-second DoE SciDAC Proposal
- SAN/OSD-Based Cluster File System
- Network Attached Storage Filer
- Hierarchical Storage Management (HSM)

Terabyte-per-second DoE SciDAC Proposal

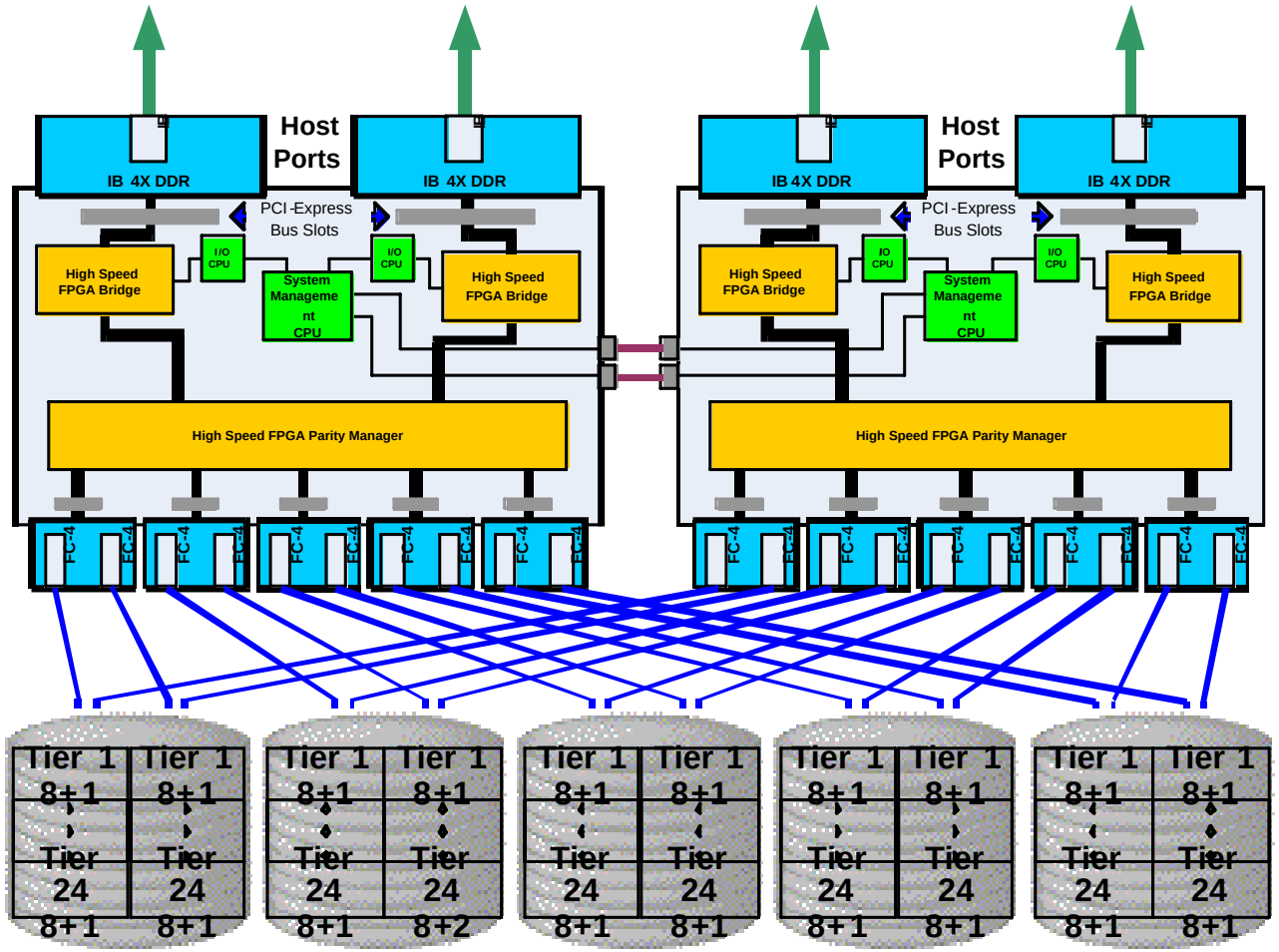
- Joint work with ORNL, Livermore, Data Direct Networks and CFS
- Goal is 1/10 Terabyte/second IO subsystem in 2008/2010 to support DoE Petaflops system
  - High-speed IB connections
  - Increased controller densities
  - 2.5-inch drives (and lots of them)
  - Lustre scaling improvements
  - Fast Cray IO node (2 Gbytes/second)
- Push technology envelope to reach this goal

Cray IO Path for 1/10 Terabyte/second



Storage Array for 1 Tbyte/Second

Building Block for 1 TB/s. System



S2A STORAGE CONTROLLER BUILDING BLOCK

8 GB/s Host Side Throughput

FC-4

STORAGE BUILDING BLOCK

Five 48 Slot Enclosure with 220 Disks

250 x Building Blocks
Delivers the Required Performance
For a 1 TB/s. System

SAN or Object-Based Cluster File System

- **Strengths:** Speed, Scalability

- **Needs Work:** Management (including interoperability, backup, recoverability), Stability, Robustness, Standards (POSIX isn't good enough), Small and fragmented commercial market

  - Interoperability is hard (multiple OS regressions, requirement of synchronous operation)
  - Stability and Robustness compete with Features and Performance: you can't have both simultaneously
    - Conservation of energy principle of engineering
  - Defragmentation of this market seems unlikely
  - Generally operating system and platform-specific solutions

Network Attached Storage Filer

- **Strengths:** Manageability, Interoperability, Moderate Scalability, Very large (and growing) commercial market

- **Needs Work:** Performance
  - Accelerated TCP/IP via TOE hardware
  - NFS RDMA (part of NFS v4 spec)
  - Global Name Space supported in hardware

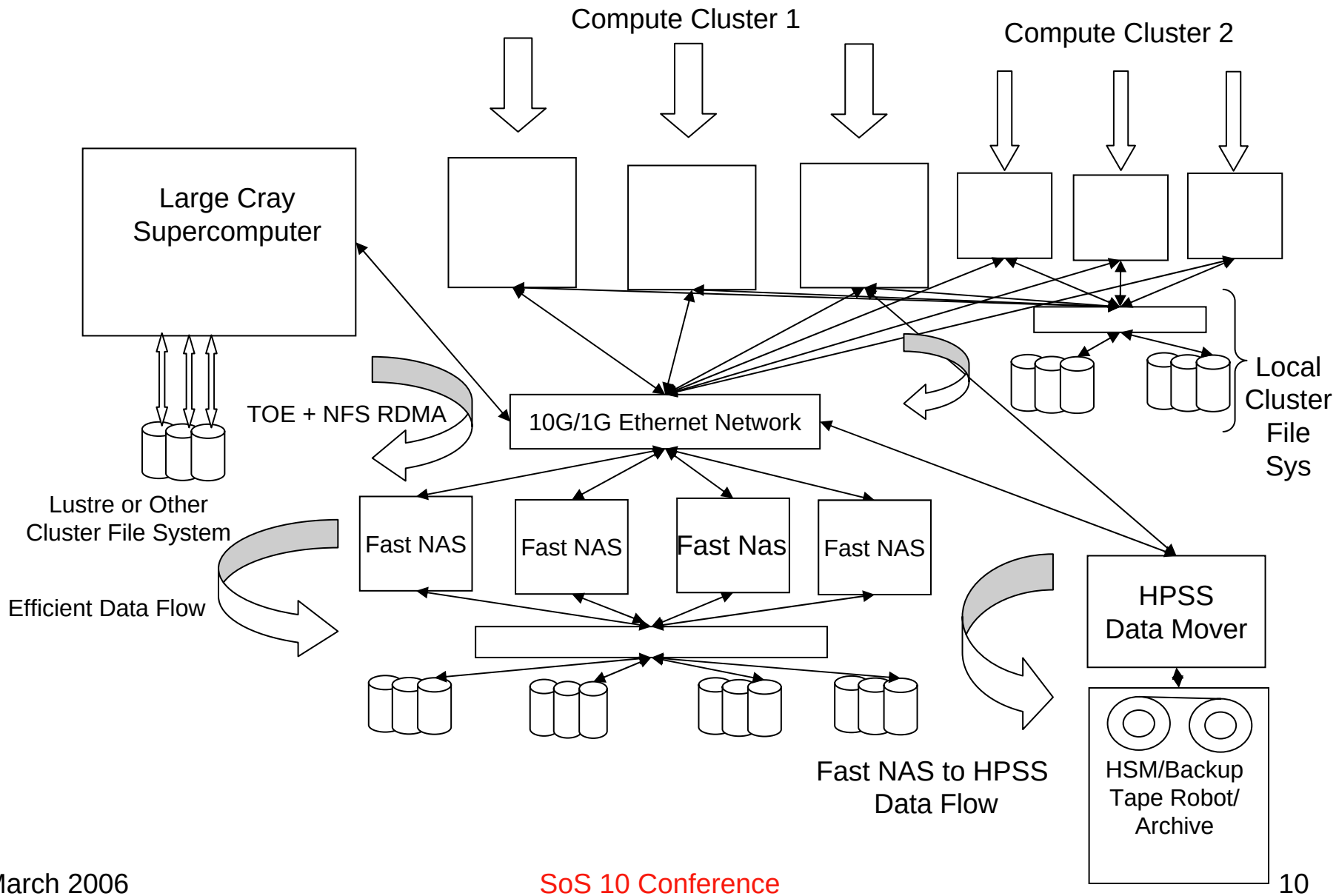
- A possible analogy:
  - Fibre Channel is high quality, fast, networked storage (Cluster File System)
  - SATA is high density, moderate speed, single node storage (Network Attached Storage)
  - You need both (right now)
  - But Tom Ruwart tells me FC is going away because 2.5-in SATA quality is high, etc. etc.

Data Archiving (HSM)

- **HSM Strengths**
  - Disk storage at the price of tape
  - Large stores, useful for sparse access patterns (1-4 Petabytes or more)
  - No real alternative to backup/archive of large HPC datasets
- **HSM Weaknesses**
  - Complex to implement and operate
  - Hard to hide fact that tape slower than disk from file system name space
  - Are commercial ILM offerings based on SATA disk an alternative?

- SATA and MAID to significantly augment tape? (speculative)
  - Simplified implementation due to reduced error cases
  - MAID much faster than tape, more closely mimics disk
  - But ILM semantics poor relative to HPC HSM semantics

A Potential Storage Architecture (CFS and NAS storage)



Cray Storage Solution

- XT3 Compute engine
  - compute nodes
  - User Login
  - XT3 local OSS
  - XT3 local MDS