Enabling Data Intensive Applications using Logistical Networking Tools

Micah Beck

Associate Professor

Logistical Computing and Internetworking Lab

Computer Science Department

University of Tennessee

DOE Network PI Meeting 9/28/2005

LoCI

LOGISTICAL COMPUTING AND INTERNETWORKING LAB

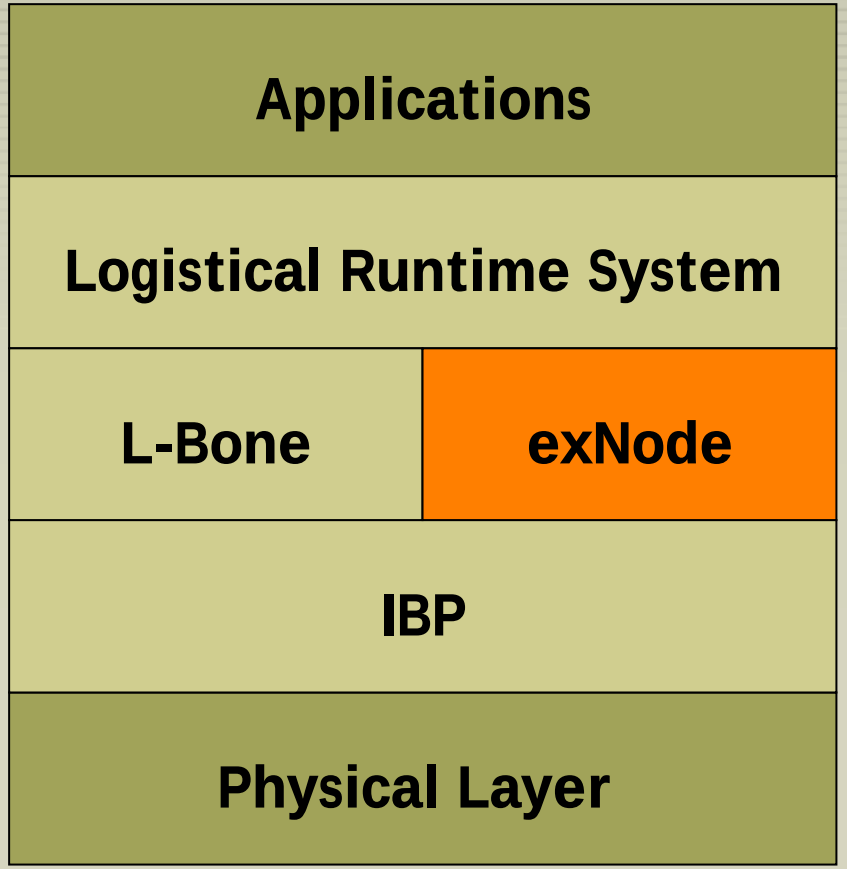
UNIVERSITY OF TENNESSEE

The Internet Backplane Protocol

- A common service for state management in a shared network
- A basis for asynchronous communication
- Scalability comes from weak assumptions:
  - Maximum size & duration of allocation
  - A highly generic, “best effort” service
  - “a weak network version of malloc”
- Robust services are built on top *in an end-to-end manner*!
- The goal is scalability analogous to the Internet

LoCI

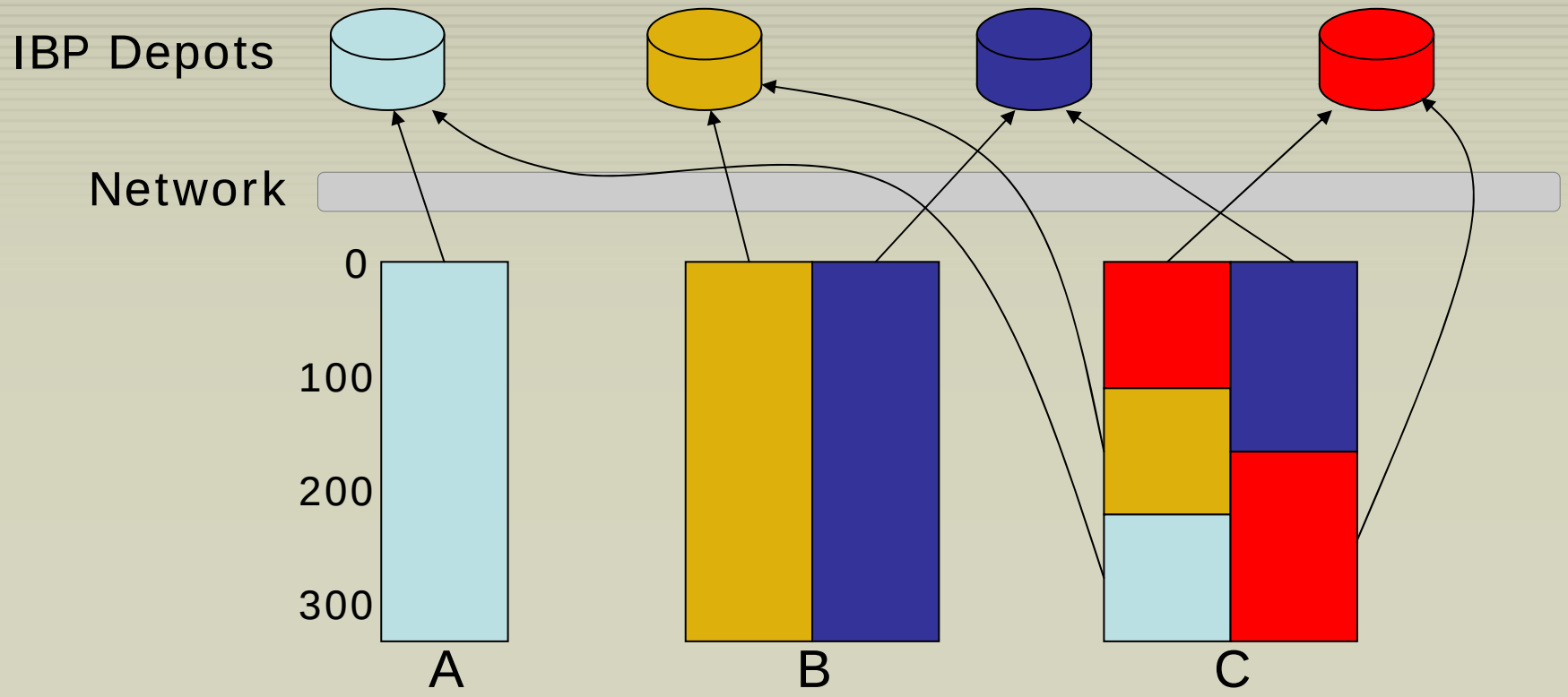
exNode



- Like Unix inode
- Provides mapping from logical view to the storage
- Allows for larger stored files than any single IBP allocation or any single IBP “depot” can provide
- Allows for replication to improve reliability and performance
- Allows unlimited annotation of the global exNode and of individual mappings
- Provides means to describe if the data was transformed before storage (e.g. encrypted)
- Serializes to XML

LoCI

Sample exNodes



LoCI

“Data in Transit”

- After being generated by an instrument or supercomputer
- Not stored in a permanent archive
- Serving the diverse purposes of a community of users and applications
- Being transferred, processed and stored to meet changing and unanticipated needs
  - Visualization
  - Data Mining
  - Collaboration
  - Distributed Computing

LoCI

SciDAC Application Impact

- **Terascale Supernova Initiative**
  (A. Mezzacappa, ONRL; J. Blondin, NCSU, D. Swesty, SUNY Stony Brook)
  - Five 1.6TB depots deployed at TSI sites
- **Energy Fusion Research** *(S. Klasky, PPPL)*
  - Depots deployed on PPPL cluster nodes
- Dataset transfers: O(1TB) @ 1-400 Mb/s
  - Simulations at NERSC and ORNL
  - Control/viz at ONRL, NCSU, Stony Brook, PPPL
  - Transfers span ESNet, Abilene
- Collaborating with two Fusion Simulation SciDAC projects to enable Logistical Networking for Data Management
- Application outreach: Combustion, Particle Physics, Earth Sci, Material Sci, …

LoCI

Porting Application I/O Libraries to Logistical Networking

- Reading and writing directly to IBP depots using LoRS functions and exNodes
  - NetCDF/L
  - HDF/L
  - libxio (POSIX functionality) – no relation to Globus XIO
  - stdio (in development)
- ROMIO Support of LN
  - Jonghyun Lee of Argonne National Laboratory
  - ADIO_LN implements abstract device
  - MPIO
  - Parallel netCDF
  - Parallel HDF5

LoCI

Indirect Management of Data

- Anecdote: VH-1 Supernova Simulation Code
  - Data distributed among 200-1000 processors
  - Highest bandwidth achieved when each processor writes a separate file independently
  - Inconvenient to manage
    - Collective I/O generates one file
    - Postprocessing combines files
- If a single entity can be described in metadata, (eg. exNode) *no data movement is required*
  - But what about architecture dependences: endianism, floating point representation?

LoCI

Goal: Flexible Management of Metadata Decoupled From Storage Resources

- Seperation of Object Storage Targets from metadata is a trend in parallel file systems
  - PVFS2, Lustre
- Object Storage Devices are becoming more abstract
  - Move away from block addressing
  - Sharing between hosts implemented at the OSD
- OSDs do not support sharing within a network community
  - Access to storage resources requires participation in distributed file system.

LoCI

Logistical and Optical Networking

- Optical switching provides ultrascale connectivity between directly attached nodes without buffering
- There are reasons why buffering may be required even when using an optical network
  - Some application communities have nodes not attached to the optical switched network
  - There may be non-homogeneity between parts of the optical network, eg in capacity, traffic
- Buffering allows connections between optical and non-optical networks
- Buffering can increase utilization of optical paths

LoCI

Whither Data Intensive Applications?

- Massive computations are part of a larger workflow that requires movement of data in the wide area
- Scientists interact with long batch computations as “slow streams” of intermediate results.
- Managing global distribution, archiving and access are seen as future challenges.
- Growing disconnect between processing and storage, systems and network
  - Part of this has to do with accounting!
- A static view of data does not model this world!

LoCI

mbeck@cs.utk.edu

http://loci.cs.utk.edu

LoCI