USC Viterbi School of Engineering

A Data Placement Service for Petascale Applications

Ann Chervenak, Robert Schuler
USC Information Sciences Institute
SciDAC Center for Enabling Distributed Petascale Science (CEDPS)

USC Viterbi School of Engineering

Motivation

- Scientific applications often perform complex computational analyses that consume and produce large data sets
  - Computational and storage resources distributed in the wide area
- The placement of data onto storage systems can have a significant impact on
  - performance of applications
  - reliability and availability of data sets
- We want to identify data placement policies that distribute data sets so that they can be
  - staged into or out of computations efficiently
  - replicated to improve performance and reliability
- This paper presents an architecture for flexible, policy-driven data placement services
  - Also implementation of first layer of this architecture

USC Viterbi School of Engineering

Outline

- Existing data placement services
- Earlier work on data placement and workflow management
- Data placement service architecture
- Implementation of first layer: lightweight bulk transfer
- Summary

Example Data Placement Service: Physics Experiment Data Export (PheDEx)

- Manages data distribution for CMS High Energy Physics Project
- High energy physics community has a hierarchical or tiered model for data distribution
  - Tier 0 at CERN: data collected, pre-processed, archived
  - Tier 1 sites: store & archive large subsets of Tier 0 data
  - Tier 2 sites: less storage, store a smaller subset of data
- Goal of PheDEx: automate data distribution processes
- PheDEx system design involves agents running at each site, communicating through a central database
- PheDex supports:
  - initial "push-based" hierarchical distribution from Tier 0 site
  - subscription-based transfer of data
  - on-demand access to data by individual sites or scientists

USC Viterbi School of Engineering

Existing Data Placement Services

- PheDEx is one example of data management services for science
  - Others include Lightweight Data Replicator (LDR) for the Laser Interferometer Gravitational Wave Observatory (LIGO) project
- Provide asynchronous data movement of large scientific data sets
  - Terabytes of data
  - Millions of files
- Disseminate subsets of data to multiple sites some time after data sets are produced
  - Based on VO policies, metadata queries, subscriptions, explicit data requests
- Stage data sets onto resources where scientists plan to run analyses

USC Viterbi School of Engineering

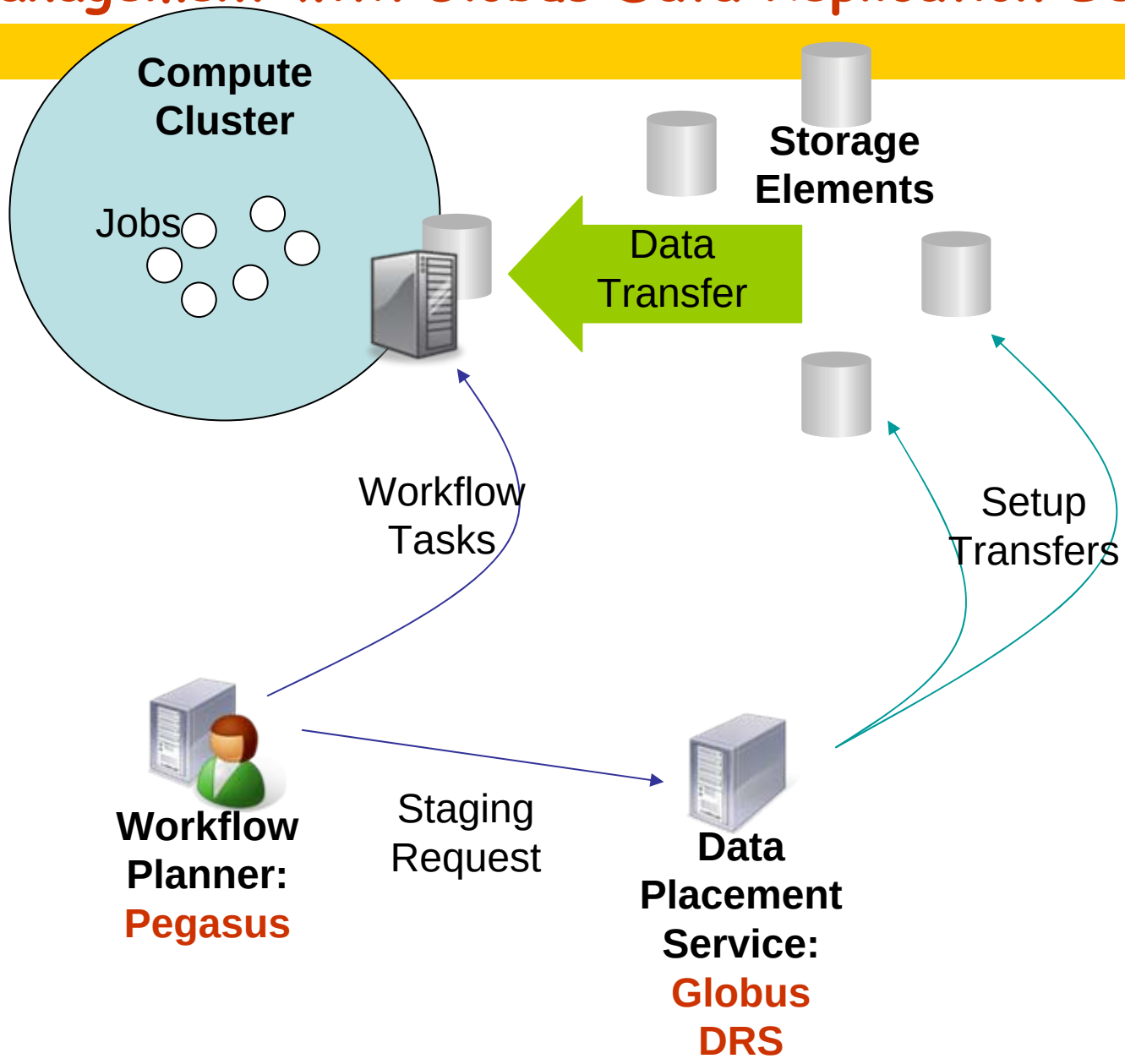
Data Placement and Workflow Management

- Studied relationship between asynchronous data placement services and workflow management systems
  - Workflow system can provide hints r.e. grouping of files, expected order of access, dependencies, etc.
- Contrasts with many existing workflow systems
  - Explicitly stage data onto computational nodes before execution
- Some explicit data staging may still be required
- Data placement has potential to
  - Significantly reduce need for on-demand data staging
  - Improve workflow execution time
- Experimental evaluation demonstrates that good placement can significantly improve workflow execution performance

"Data Placement for Scientific Applications in Distributed Environments," Ann Chervenak, Ewa Deelman, Miron Livny, Mei-Hui Su, Rob Schuler, Shishir Bharathi, Gaurang Mehta, Karan Vahi, in Proceedings of Grid 2007 Conference, Austin, TX, September 2007.

USC Viterbi School of Engineering

Approach: Combine Pegasus Workflow Management with Globus Data Replication Service



USC Viterbi School of Engineering

Data Placement Service

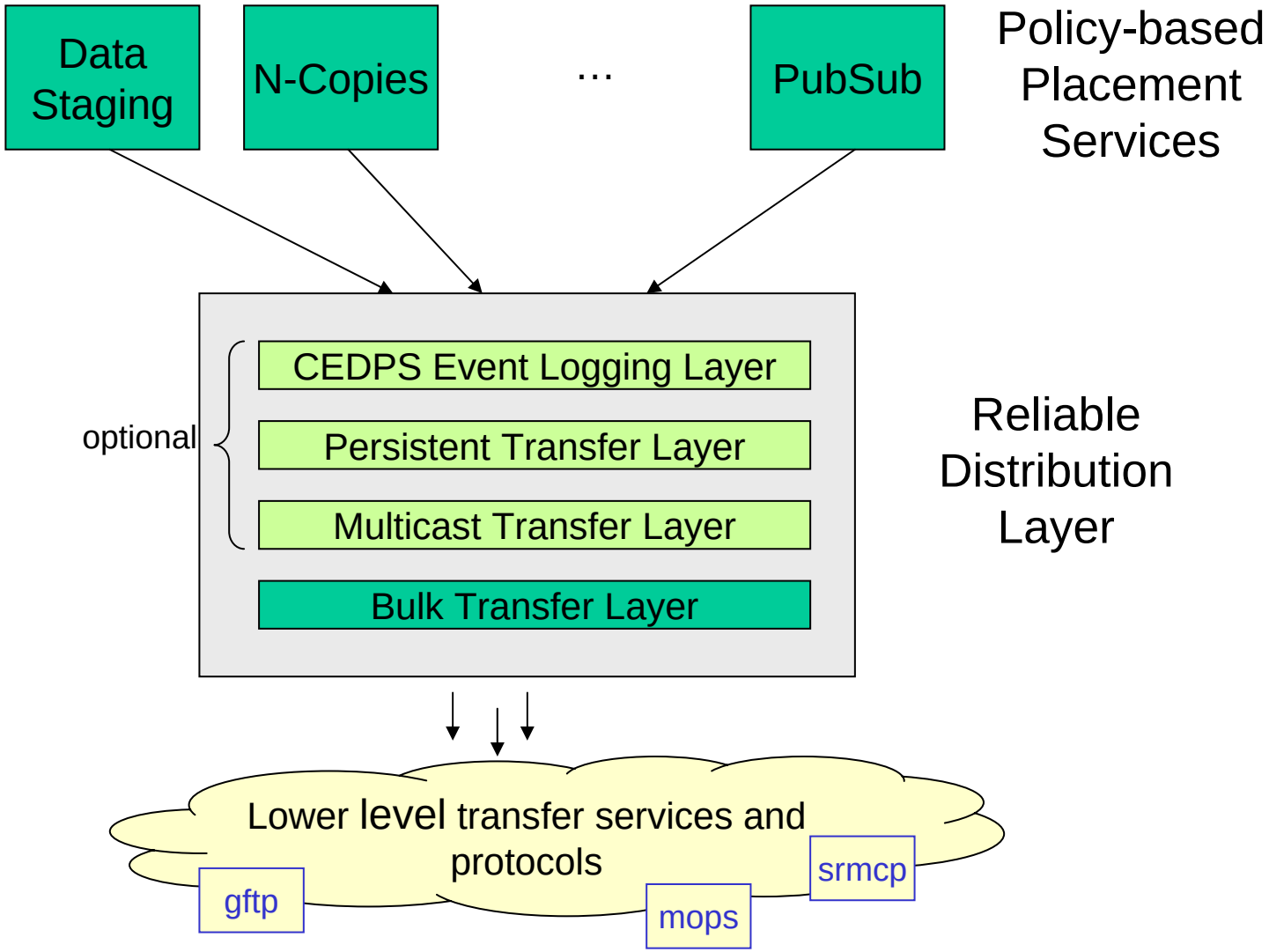
- Want to build a general data placement service that can be used by a variety of application communities

- DPS is a layered architecture made up of:
  - Policy-based services
    - Data staging, N-copies, publish/subscribe, push/pull, etc.
  - Reliable distribution layer
    - Logging, persistence, "multicast" style transfer, bulk transfer management

- The *Bulk Transfer Service* is the first layer implemented in the Reliable Distribution Layer stack

USC Viterbi School of Engineering

Data Placement Service



USC Viterbi School of Engineering

Bulk Transfer Service (BuTrS)

A *lightweight*, *simple*, *efficient* utility for bulk data transfer with priorities

- Lightweight
  - Single .jar file
  - Minimal 3rd-party dependencies
- Simple
  - No admininstration overhead
  - Basic Java API
- Efficient
  - Concurrent transfers on all available pair-wise links
  - In-memory, efficient data structures for transfer queues

- Release notes:
  - http://www.cedps.net/wiki/index.php/Dps10ReleaseNotes

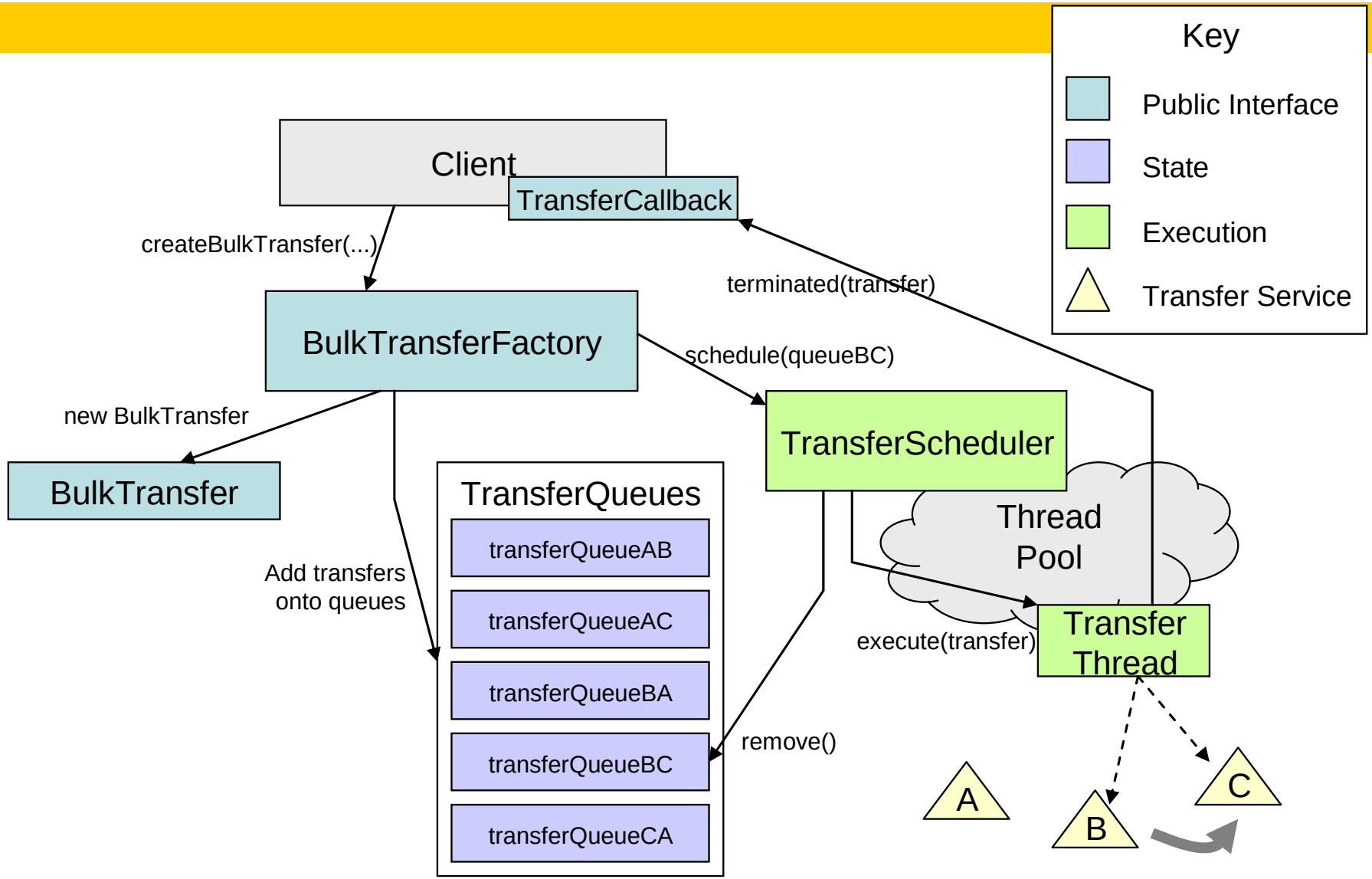
USC Viterbi School of Engineering

BuTrS Features

- Create and destroy bulk transfers
- Change priority of individual data transfer
  - Needed by clients of data placement service
- Find data transfer object by
  - (source, sink) URL pair
  - Transfer status
- Callback notification of transfer status change
- Supports GridFTP third-party transfers and most GridFTP transfer options
- Retries failed transfers up to defined maximum

USC Viterbi School of Engineering

Bulk Transfer Service



USC Viterbi School of Engineering

Ongoing and Future Work

- Continued measurements and simulations of interaction between workflow management systems and data placement services
  - Efficient stage-in and stage-out of data sets for workflows
  - Replicating data for performance and reliability
- Ongoing development
  - Bulk transfer implementation, release October 2007
  - Building additional layers for reliable distribution and policy-driven placement
- Seeking application partners to drive data placement policies

USC Viterbi School of Engineering

Acknowledgements

This work is supported by:

- The Department of Energy's Scientific Discovery through Advanced Computing II program under grant DE-FC02-06ER25757 (Center for Enabling Distributed Petascale Science CEDPS)

- CEDPS Participants:
  - USC Information Science Institute
  - Argonne National Laboratory
  - Lawrence Berkeley National Laboratory
  - University of Wisconsin Madison
  - Fermi National Laboratory