Argonne NATIONAL LABORATORY

# Concerted Flows: Infrastructure for Terabit/s Data Transfer

**Venkat Vishwanath**

venkat@anl.gov

Raj Kettimuthu, Eunsung Jung+, Jun Yi*,
Steve Tuecke, Mark Hereld, Mike Papka,
Bob Grossman and Ian Foster

Argonne National Laboratory

U.S. DEPARTMENT OF ENERGY

# Concerted Flows: Infrastructure for Terabit/s Data Transfer

Argonne NATIONAL LABORATORY

## Objectives

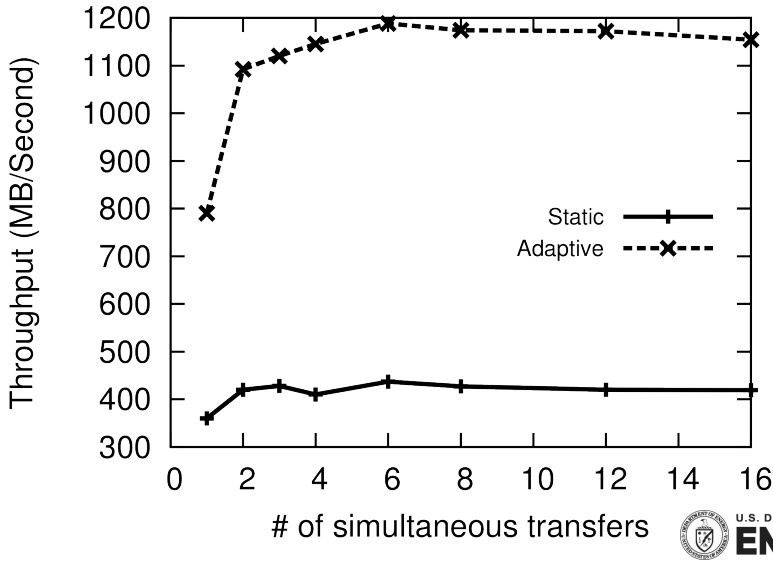
- Develop new parallel protocols that scale to Terabit/s networks
- Capture the diverse flow characteristics and needs of applications
- Create data transfer benchmark kernels for representative applications
- Exploit parallel end system topology

## Impact

- Enable DOE applications to make effective use of future end systems and advanced network infrastructure
- Build knowledge base capturing the data transfer patterns of several DOE applications

## Progress and Accomplishments

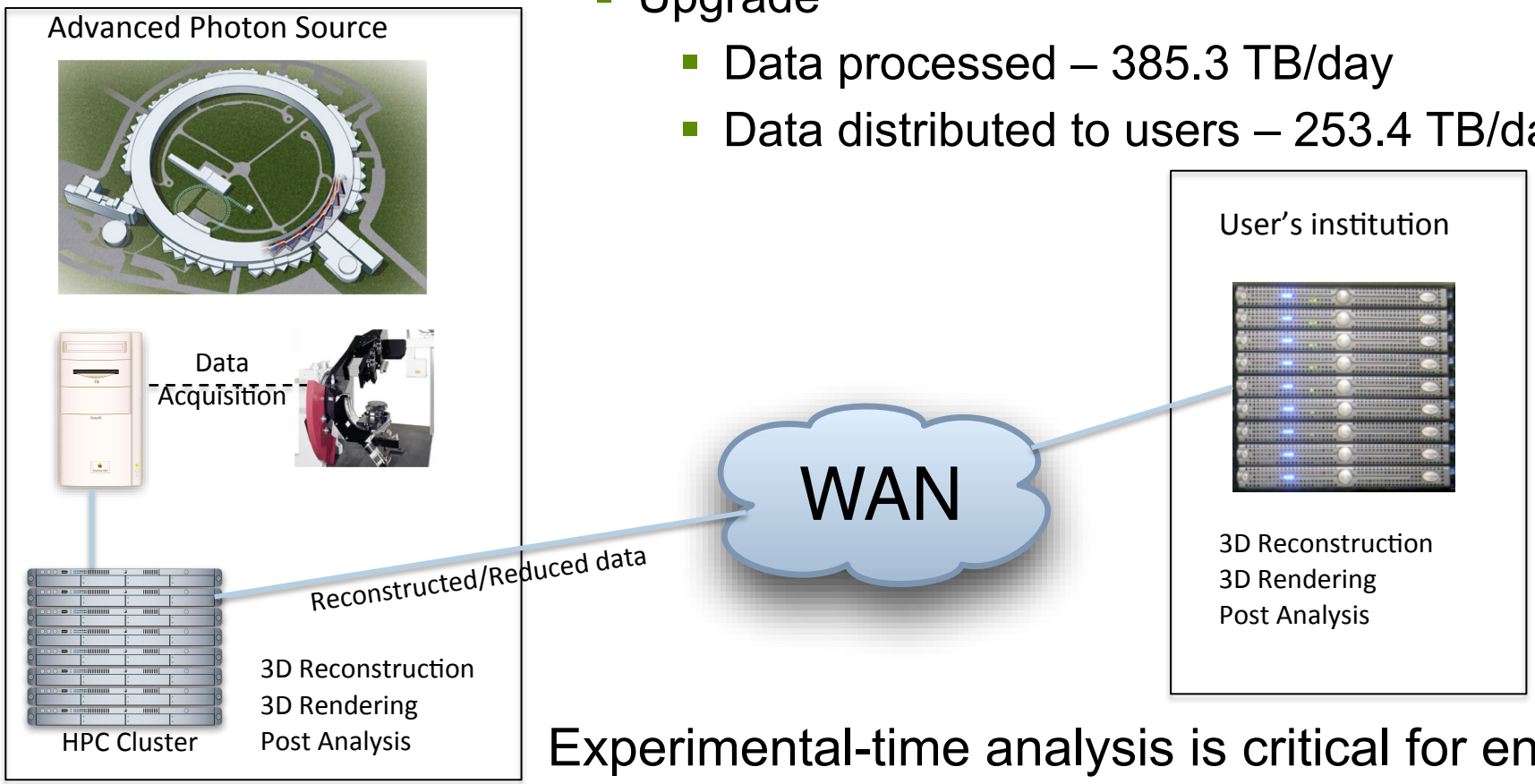
- Application kernel for tomography workflow at Advanced Photon Source
- Abstraction to evaluate multiple protocols
- Developed a prototype end system aware and network aware parallel data movement
- End system aware data movement achieves up to 3x improvement in throughput




U.S. DEPARTMENT OF ENERGY

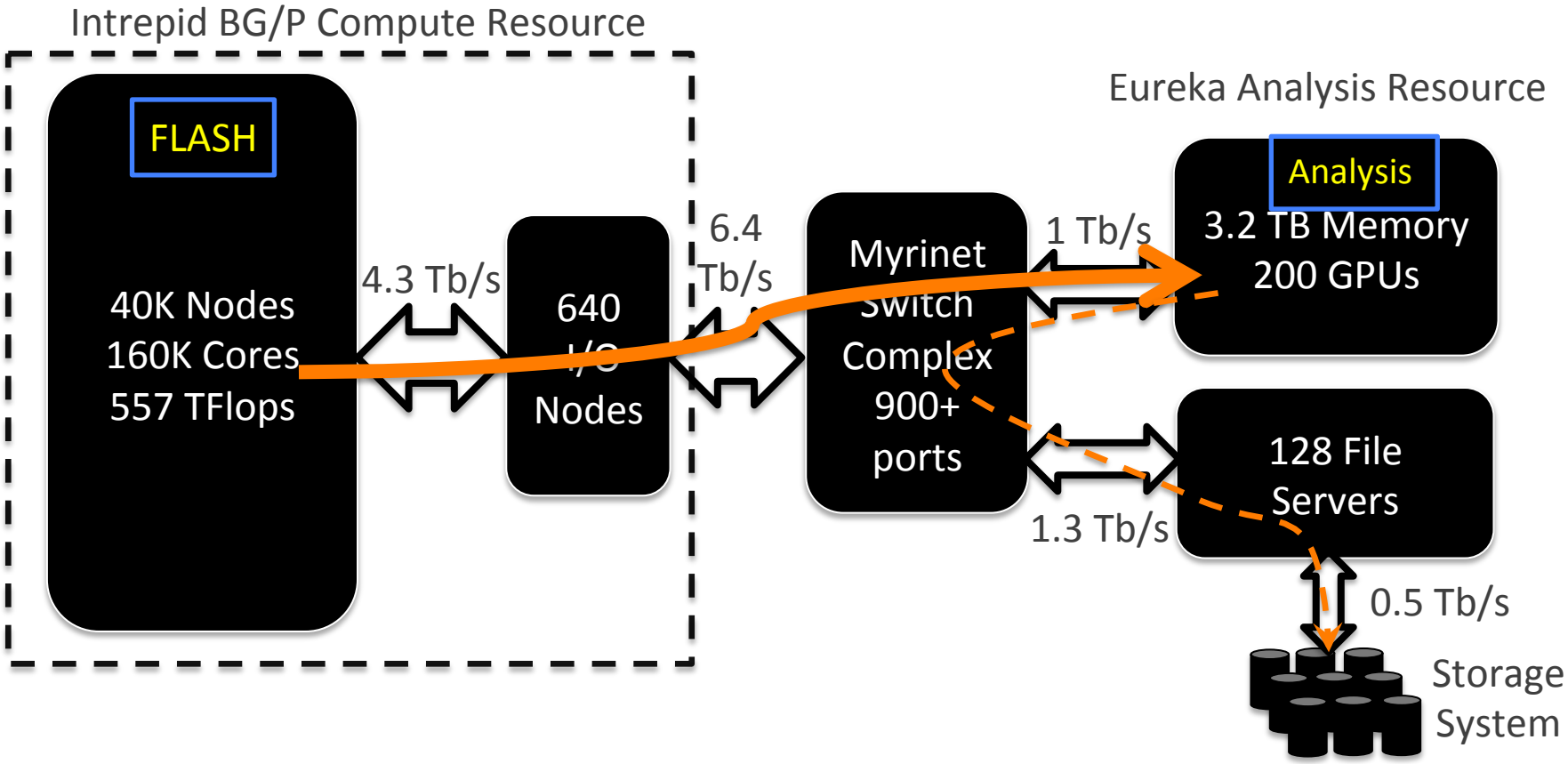
# Tomography at APS

- Current
  - Data processed – 5.6 TB/day
  - Data distributed to users – 3.3 TB/day
- Upgrade
  - Data processed – 385.3 TB/day
  - Data distributed to users – 253.4 TB/day



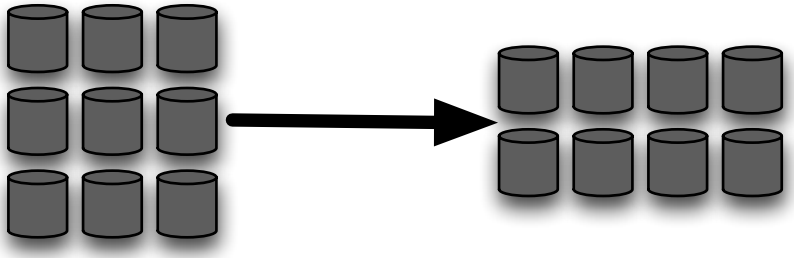

Experimental-time analysis is critical for enabling interactive changes to experiment parameters

# Simulation-time Data Staging, Analysis and Visualization of FLASH Astrophysics Simulation

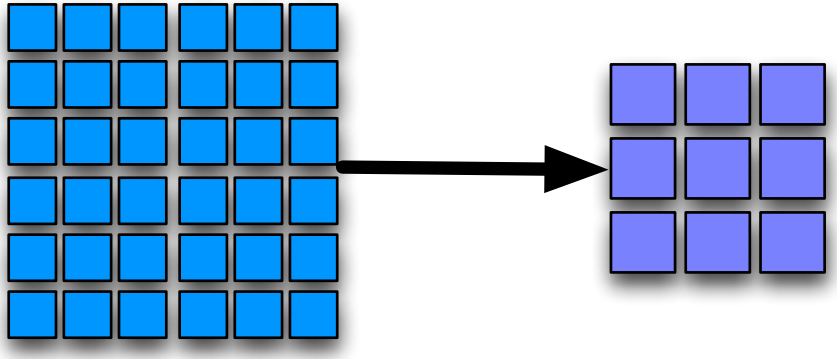



Simulation-time data analysis is critical to reduce the data written to storage and to generate faster insights. This is of critical importance for doing productive science at Petascale and beyond.

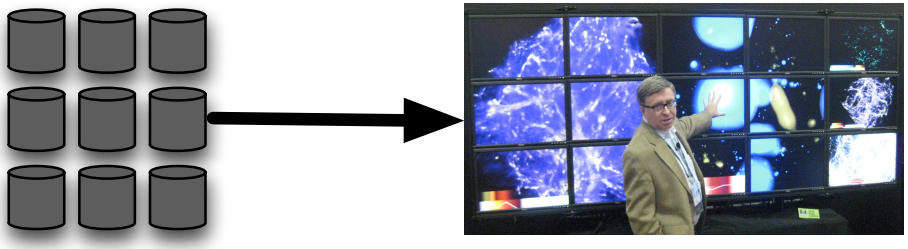
# Data Movement Trends



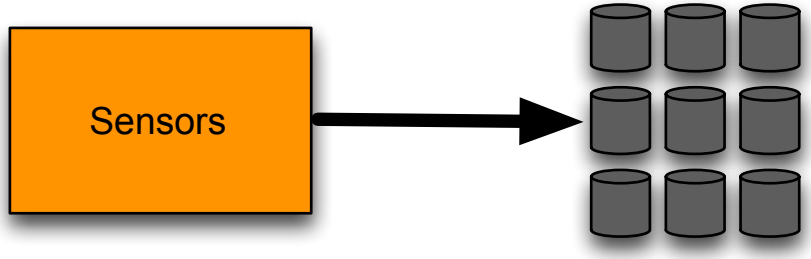
Disk-to-Disk Transfers



Memory-to-Memory Transfers



Disk-to-Memory Transfers




Memory-to-Disk Transfers

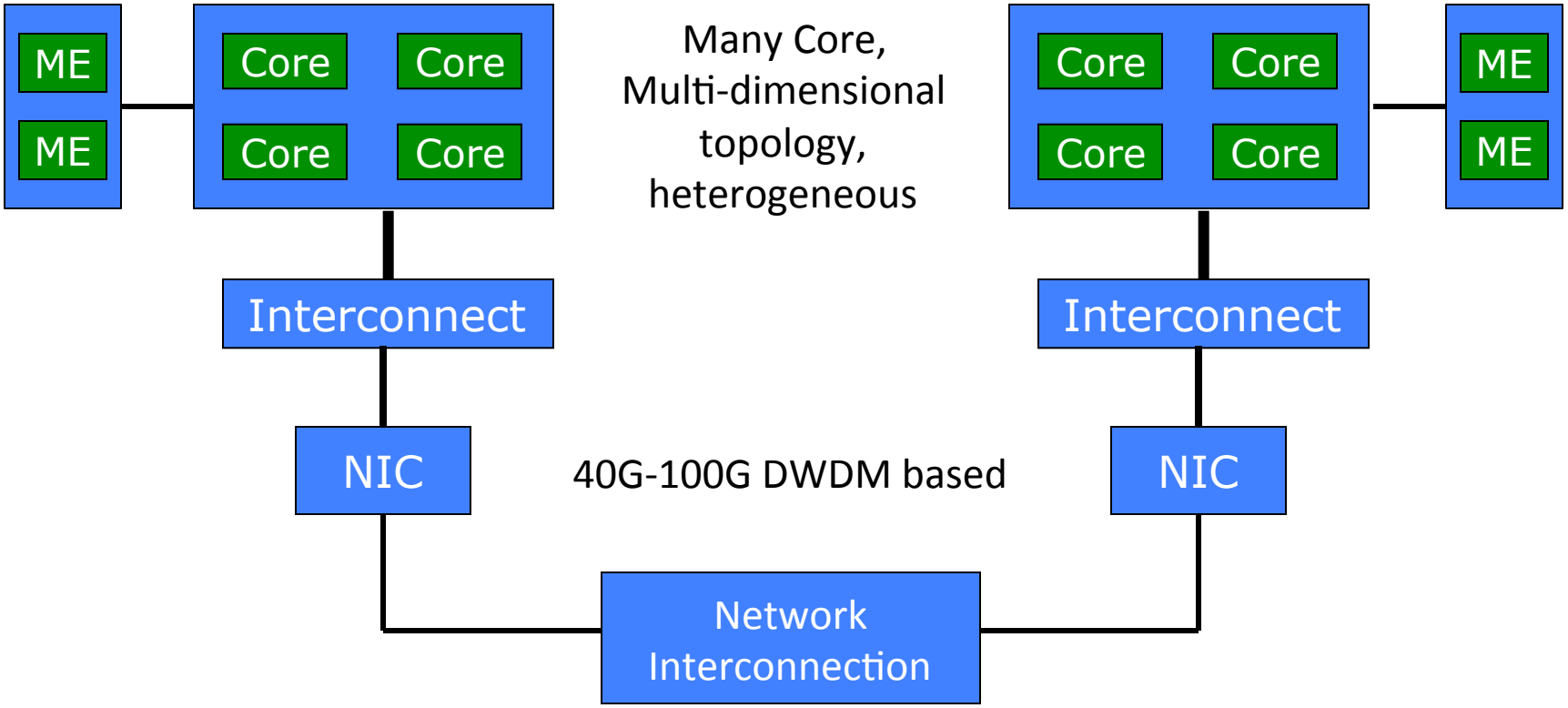
Data Movement is being increasingly characterized by Parallel M-to-N Data Flows

# Characteristics of Application Flows

| App | Type of Flow | # of Flows | BW | Latency | Burstiness | Size | Protocol |
|---|---|---|---|---|---|---|---|
| Globus Online | Data | 1 per node | High | N | Y | Large | TCP, UDT |
| | Control | 1 per session | Low | Y | Y | Small | TCP |
| APS | Data | 1 per detector | High | N | Y | Large | TCP |
| | Control | 1 per app | Low | Y | Y | Small | TCP |
| FLASH Simulation-time Analysis | Data | 1 per core | High | N* | Y | Variable | TCP, RDMA |
| | Control | 1 per app | Low | Y | y | Small | TCP, RDMA |
| ENZO Remote Viz | Data | 1 per display | High | Y | N | Large | TCP, UDP |
| | Control | 1 per app | Low | Y | Y | Small | TCP |

A mechanism to characterize and model an application's data movement behavior will be critical to better architect future networks

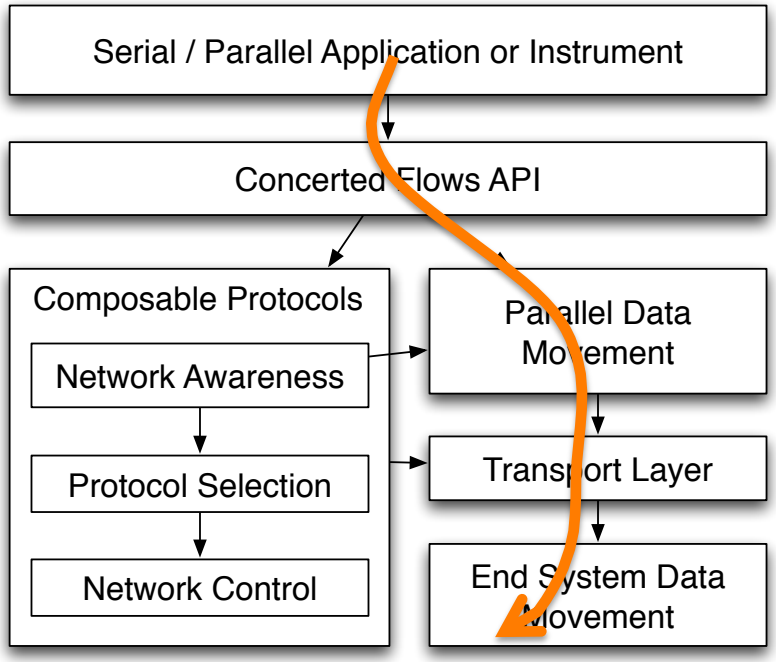
# End Systems



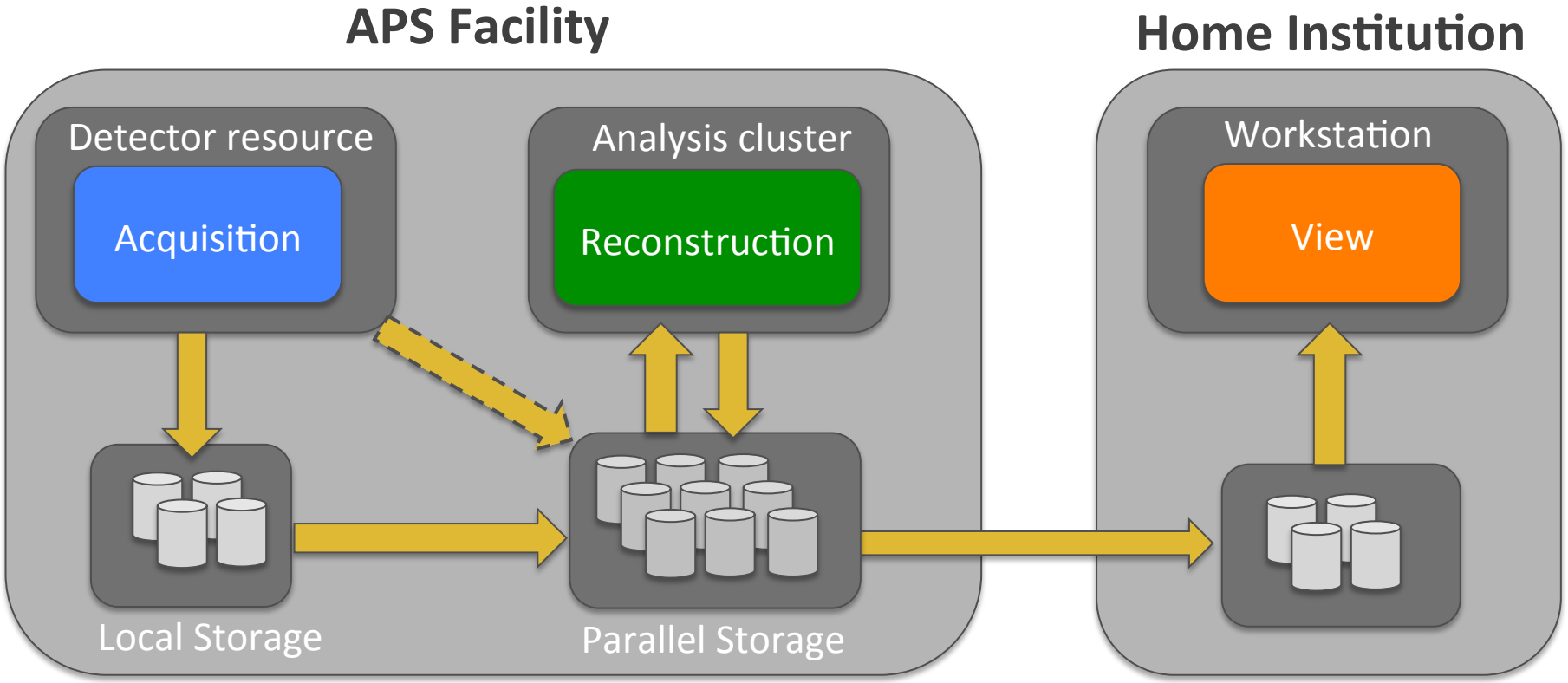

Applications need to contend with the deep and complex system hierarchies and take advantage of parallelism in the various sub-systems

# Objectives

- Develop concerted flows API
  - Capture the requirements of the application
  - Capture the characteristics of various components in the end-to-end path
    - Network, End-systems
- Create data transfer kernels for representative applications
  - Flash, Enzo, Select APS beamlines, Globus Online
- Benchmarking utilities for concerted flows (M-to-N)
- Performance optimizations at end systems and in a LAN



# APS Tomography Data Workflow




| | 2013 | 2015 |
|---|---|---|
| Beamline A | 128- 400 MB/s | 1-3 GB/s |
| Beamline B | 4- 400 MB/s | 8 GB/s |

Realtime analysis and reconstruction is of increasing importance for APS

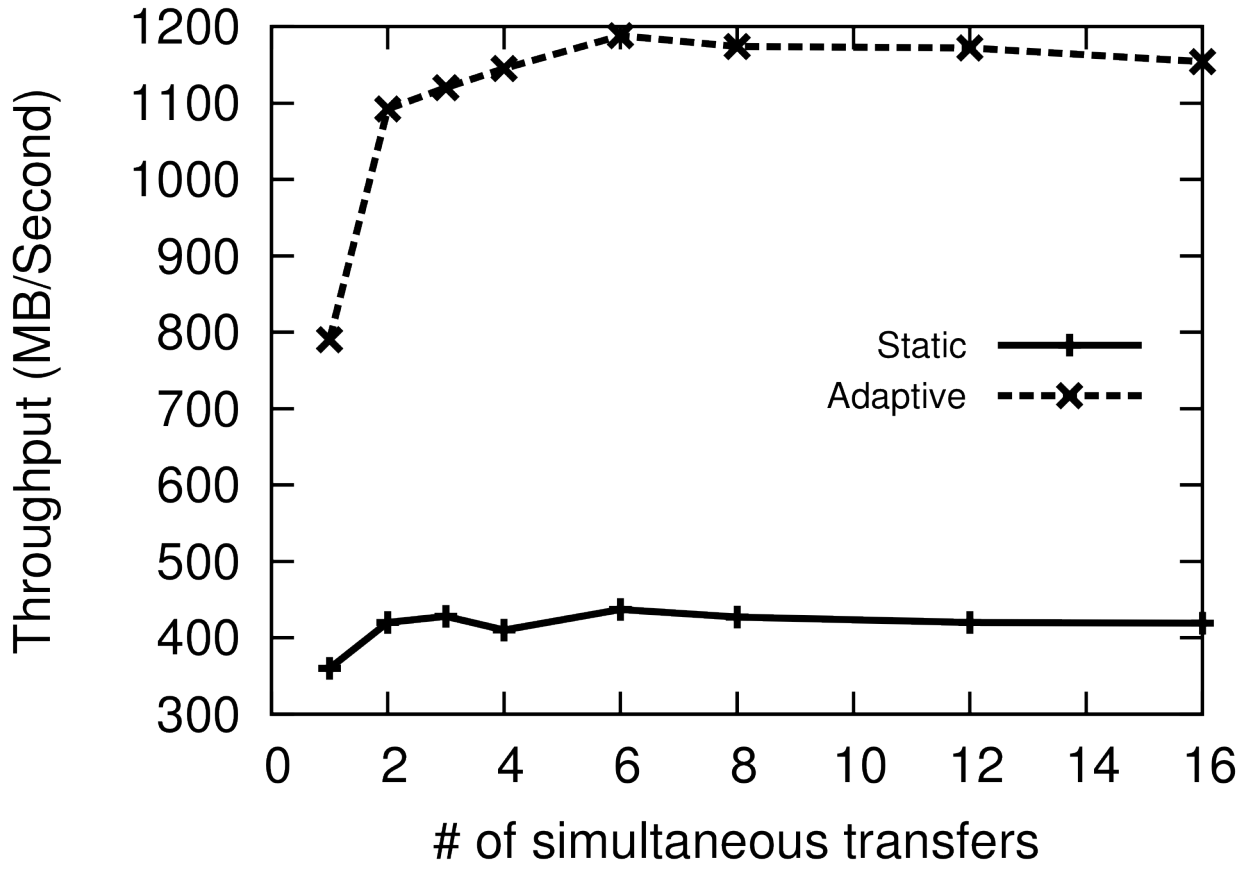
# Modeling Application Behavior using Code Skeletons

## Language features:

- Control flow
  - Parallelism
  - Reduction
  - Loops
  - Function calls
- Data flow
  - Arrays and data types
  - Indices in accesses
  - Allocation / Deallocation
- Communication
  - Point-to-point
  - Collective operations
- Characteristics
  - Instruction mix
  - Computation intensity

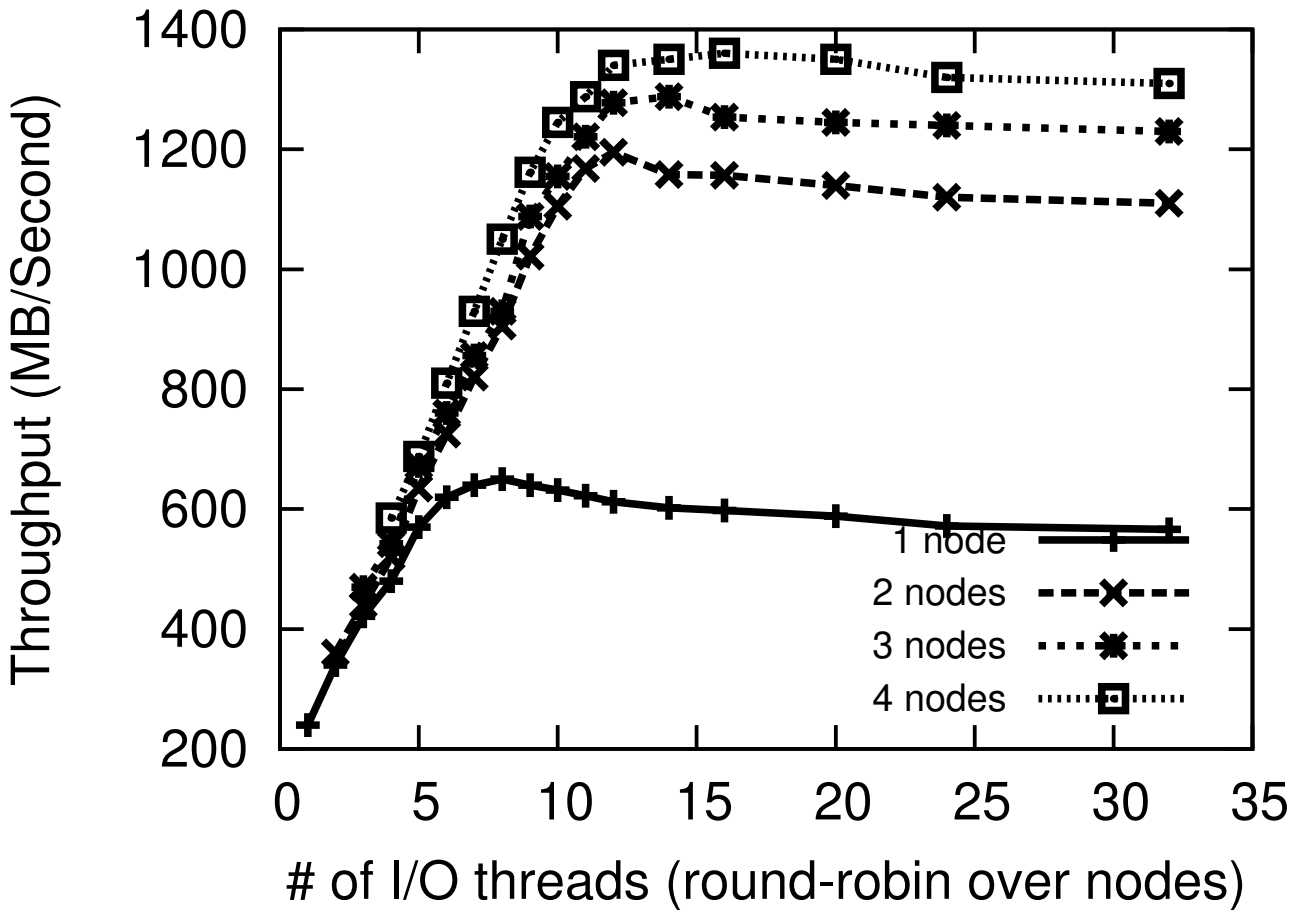
Investigating using code skeletons and swift to model the data movement characteristics of applications

# Resource Aware Data Movement



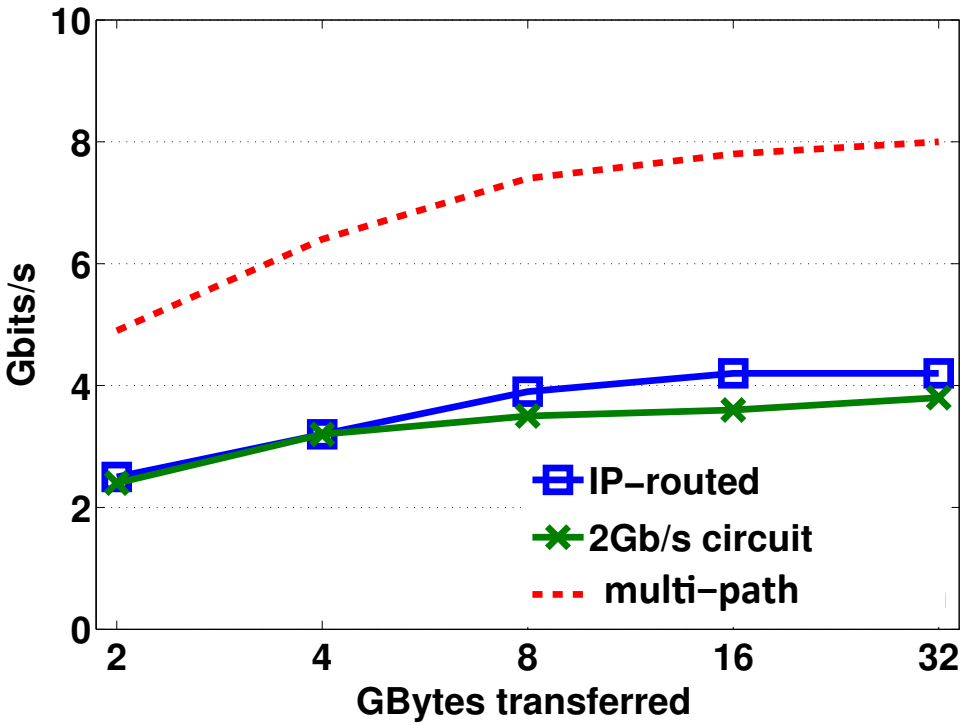

Disk-to-Disk throughput over heterogeneous nodes (combination of 1G and 10G on 4 nodes) between Argonne and NERSC using resource aware data movement has **300% improvement** over current GridFTP mechanisms

# Optimizing for Storage Systems

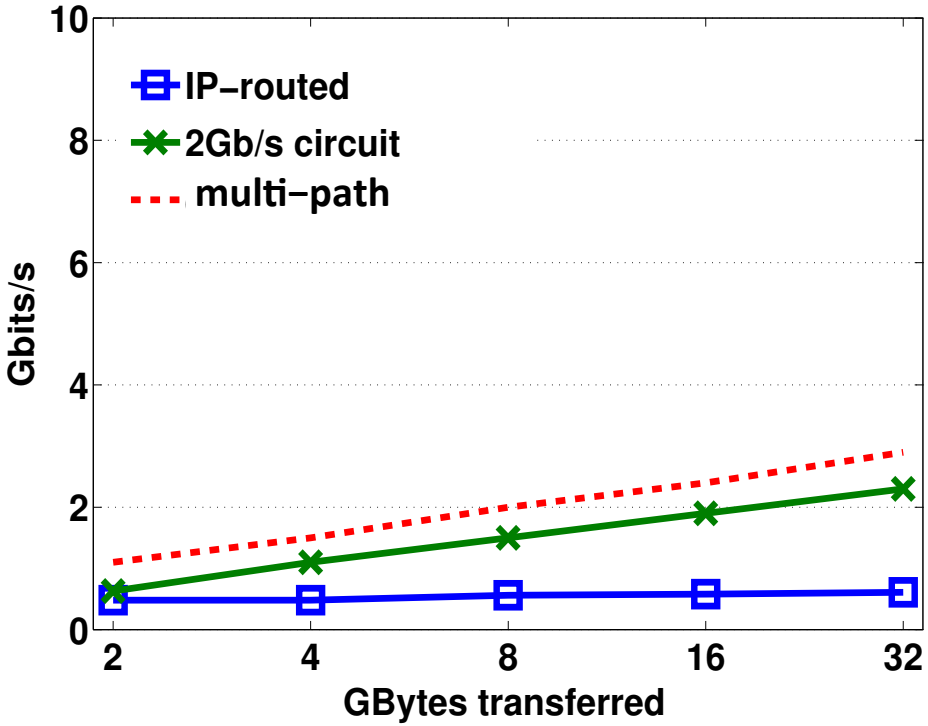



Exploiting parallelism using multiple I/O and network threads is critical for improving end-to-end transfer time

# Improving Data Movement: Multipathing




ANL and NERSC




UMichigan and Caltech

Exploiting both dedicated and best-effort paths improves the achievable throughput

# Publications

- “Toward Characterization of Data Movement in Large-Scale Scientific Applications”, Jun Yi, Rajkumar Kettimuthu, and Venkatram Vishwanath, 8th IEEE International Conference on eScience, Oct. 2012.
- "Exploiting Network Parallelism for Improving Data Transfer Performance", Daniel Gunter, Rajkumar Kettimuthu, Ezra Kissel, Martin Swany and Jason Zurawski, IEEE/ACM Annual SuperComputing Conference (SC12) Companion Volume, Nov 2012.
- "Accelerating Data Movement Leveraging Endsystem and Network Parallelism", Jun Yi, Raj Kettimuthu, Venkatram Vishwanath, IEEE/ACM Annual SuperComputing Conference (SC12) Worskhop on Network-aware Data Management

# M-to-N Data Movement Demo

- Parallel Memory-to-Memory Data Movement
- 24 Nodes of the ALCF Eureka Cluster
  - Each Node has a 10G Myrinet Network Interface
- Parallel Data Movement leverages XIO enabling one to test various protocols by swapping drivers
  - Current demo will use the TCP drivers
- Data Movement exploits parallelism within a node by using multiple threads

# Questions