Streamlining Computation and Communication for Distributed Science Workflows

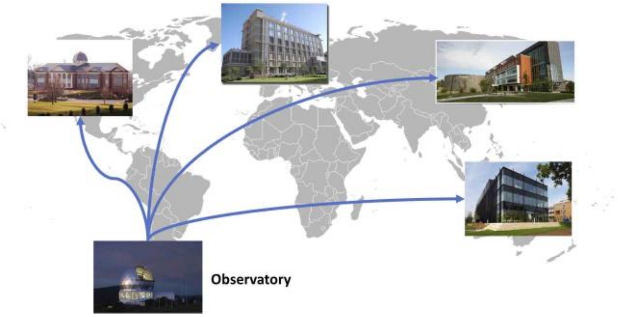
Engin Arslan

University of Nevada Reno

University of Nevada, Reno

Distributed Science

- Remote data processing
  - Data is moved from data collection facilities to HPC clusters to analyze



- Collaborative projects
  - Data sharing to enable collaboration

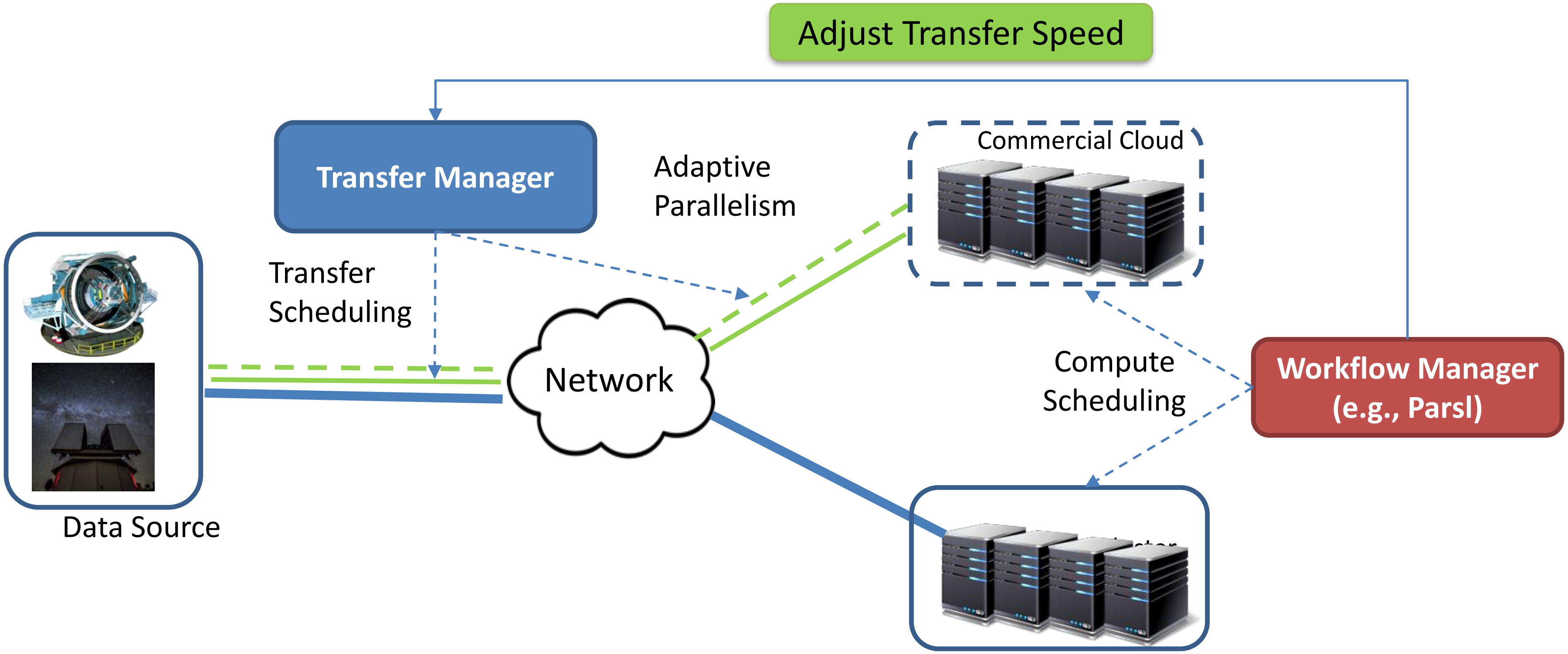
- Reproducible research
  - Central repositories to store data

NCBI

How to process large-scale remote data?

1. Download first, process later
   - 👎 Long execution time → Transfer time + Processing time
   - 👎 Need for large staging space → Potentially increased storage cost
2. Streamline compute and transfer
   - 👎 Mismatch of compute and transfer speeds
   - 👎 Performance variability due to interference

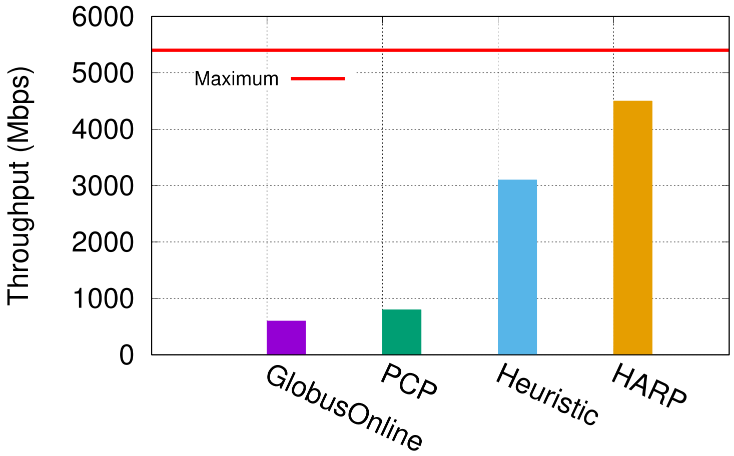
End-to-End Workflow Parallelism

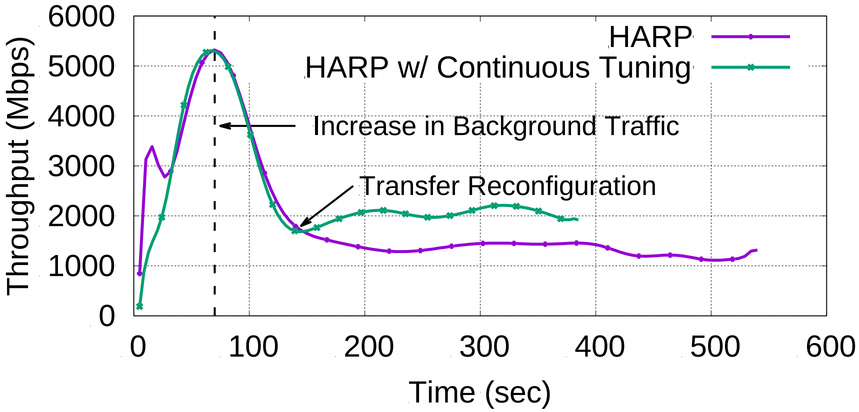


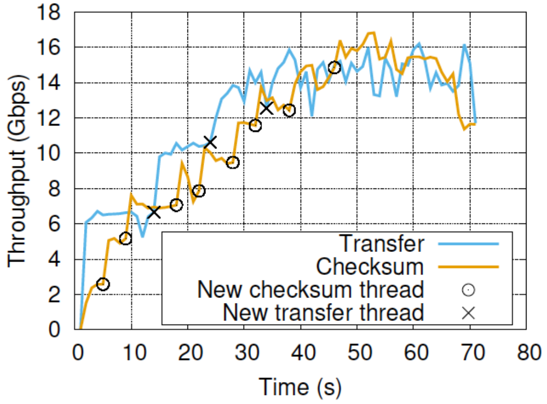
Real-time Data Transfer Tuning with HARP

High-performance data transfers

Responsive to changing conditions

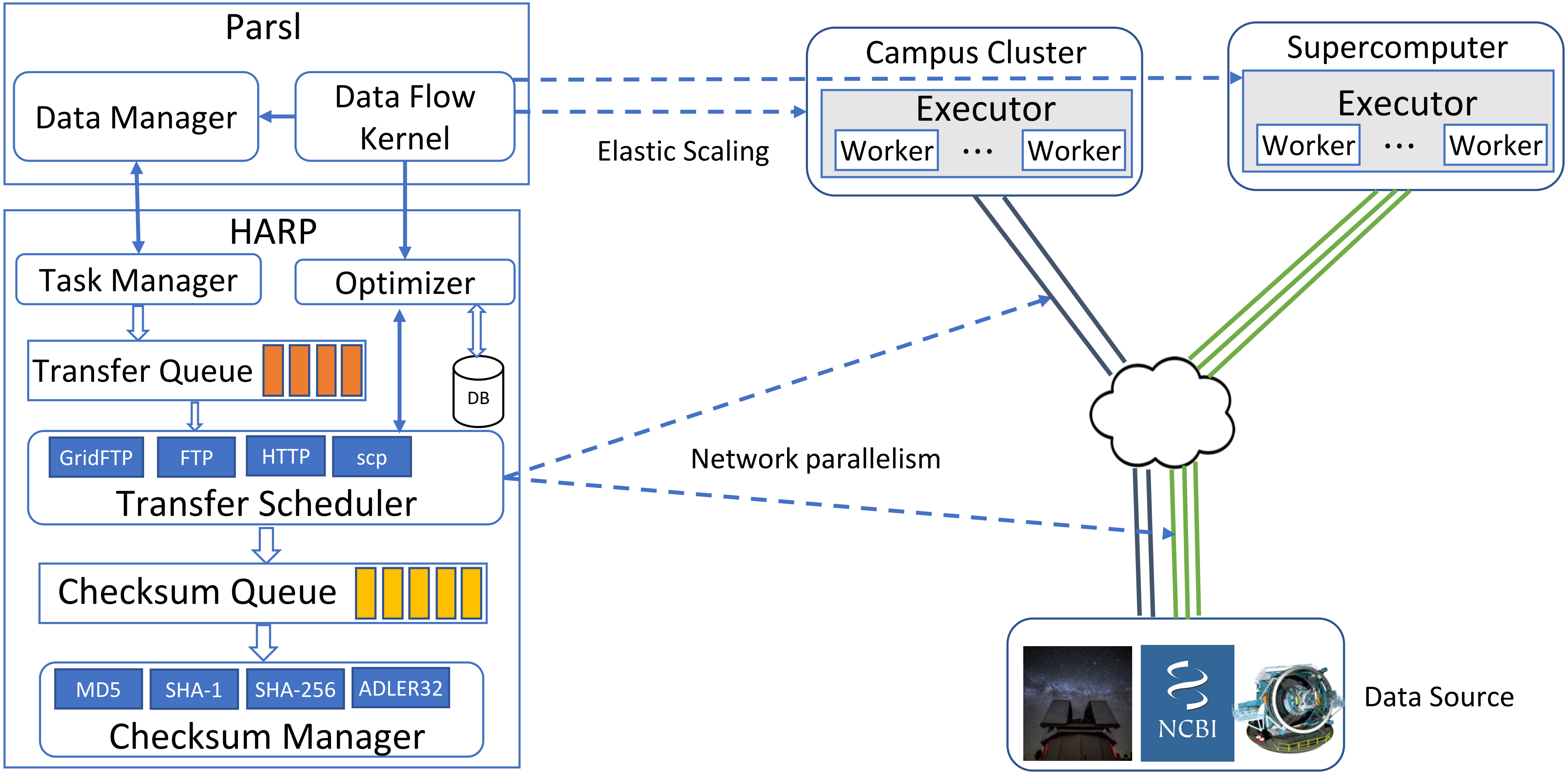






Scalable Integrity Verification

Parsl Integration



Why Parsl?

- Supports computational parallelism
- Easy to integrate new data transfer module
- Responsive and helpful development team

Thanks!