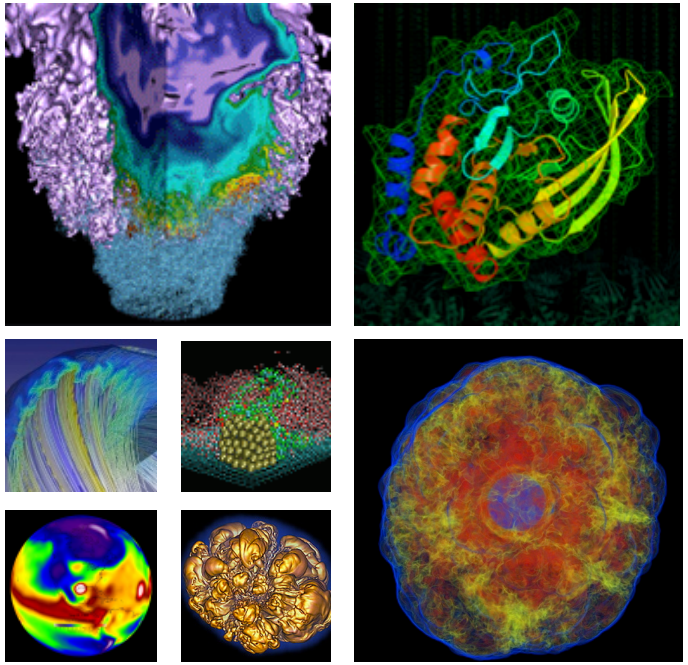
Characterizing burst buffers at extreme scale using the TOKIO framework



Glenn K. Lockwood

Advanced Technologies Group

November 16, 2016

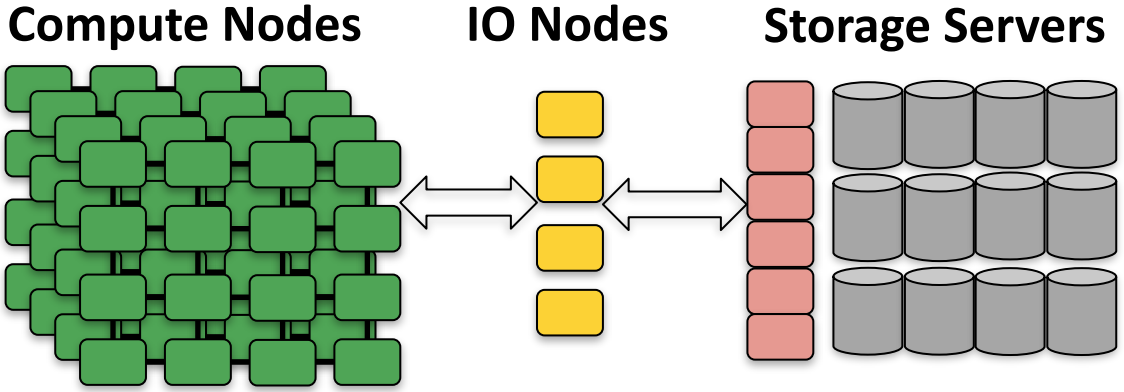
I/O at NERSC: Extreme-Scale + Diverse

- **Big file systems**
  - 35 PB Lustre
  - 10 PB GPFS
  - 1.8 PB DataWarp burst buffer
- **Lots of data movement**
  - 440 TiB/day read
  - 400 TiB/day written
- **Very diverse sources of I/O**
  - 5,576 + 11,988 compute nodes
  - 15,000 jobs/day
  - 4.1 million CPU hrs/day
  - all job sizes

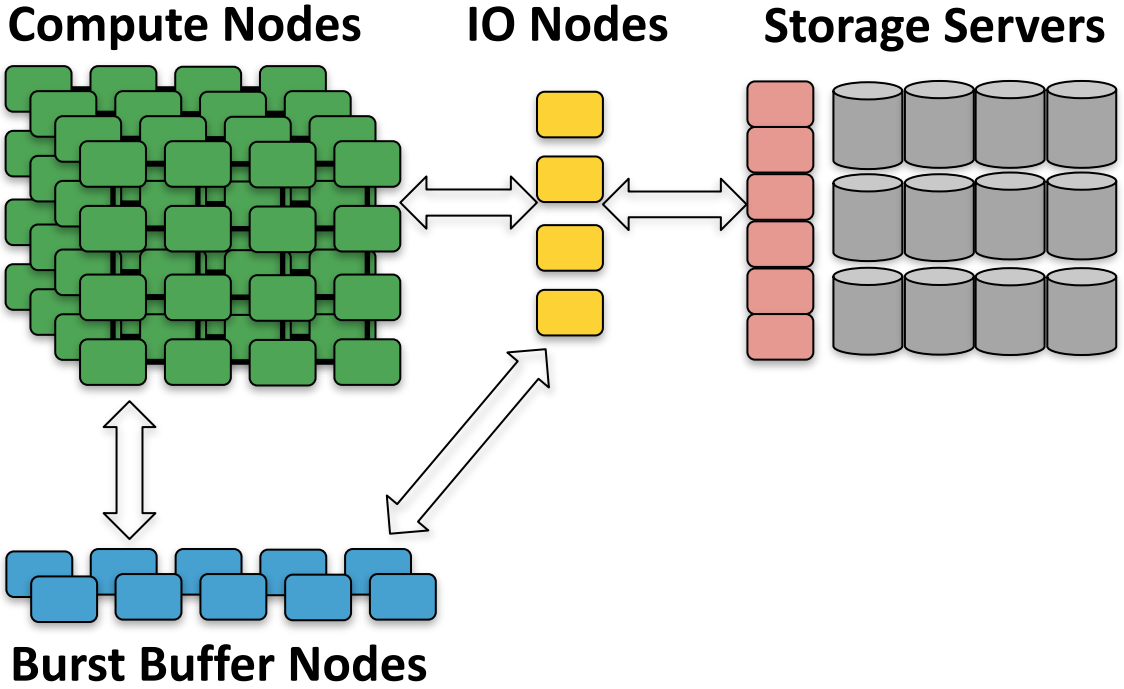
I/O at NERSC: Extreme-Scale + Diverse

- **Big file systems**
  - 35 PB Lustre
  - 10 PB GPFS
  - 1.8 PB DataWarp burst buffer
- **Lots of data movement**
  - 440 TiB/day read
  - 400 TiB/day written
- **Very diverse sources of I/O**
  - 5,576 + 11,988 compute nodes
  - 15,000 jobs/day
  - 4.1 million CPU hrs/day
  - all job sizes

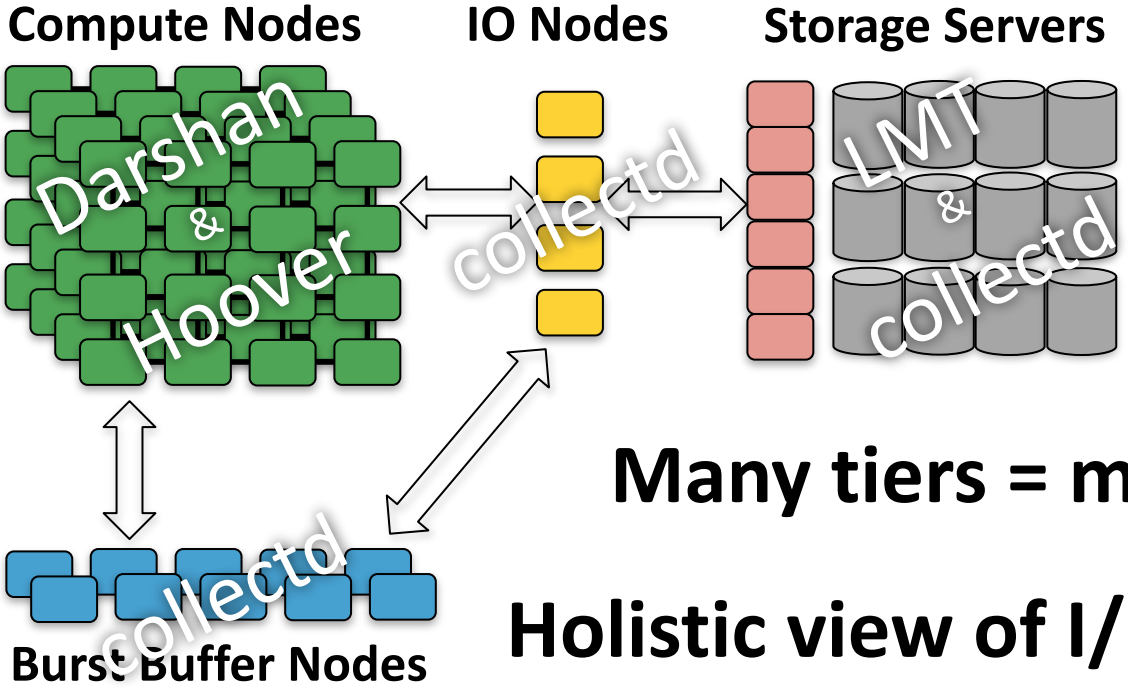
Burst Buffers Complicate Life (for those of us in the room)



Burst Buffers Complicate Life (for those of us in the room)



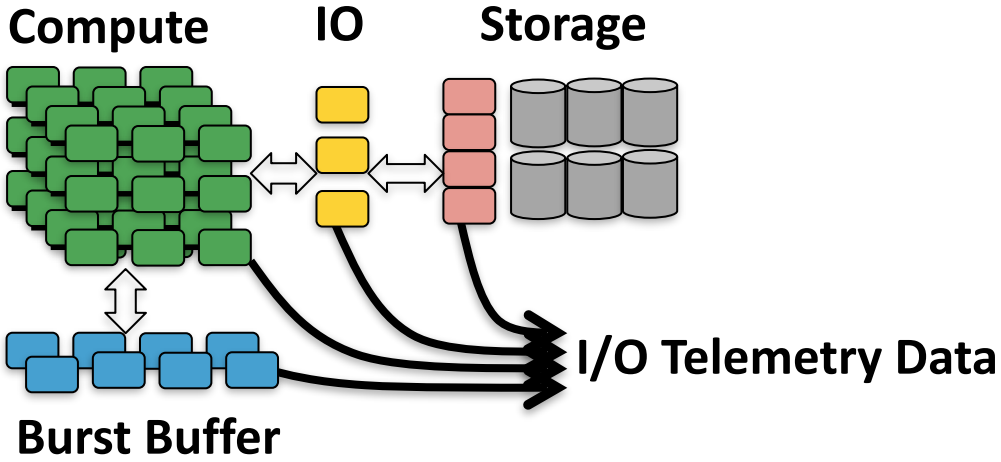
Burst Buffers Complicate Life (for those of us in the room)



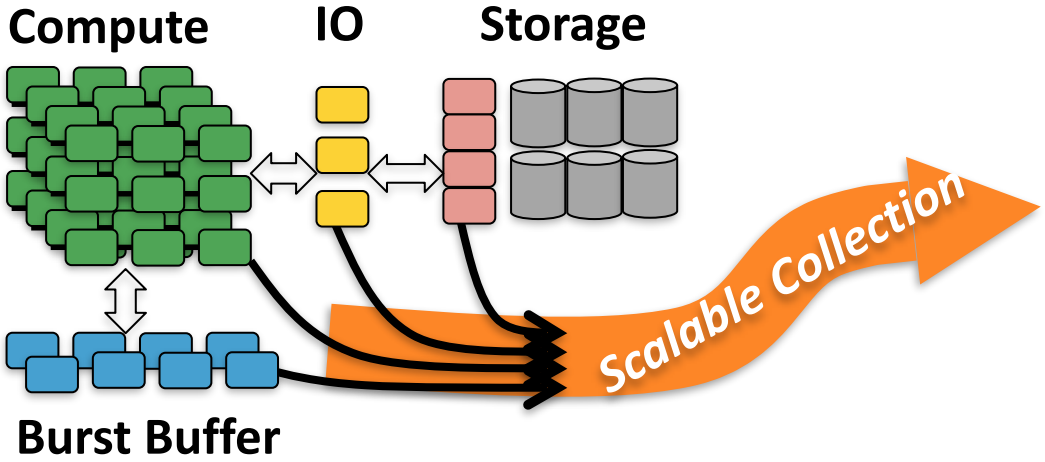
Many tiers = many tools!

Holistic view of I/O is *essential*

TOKIO: TOtal Knowledge of I/O

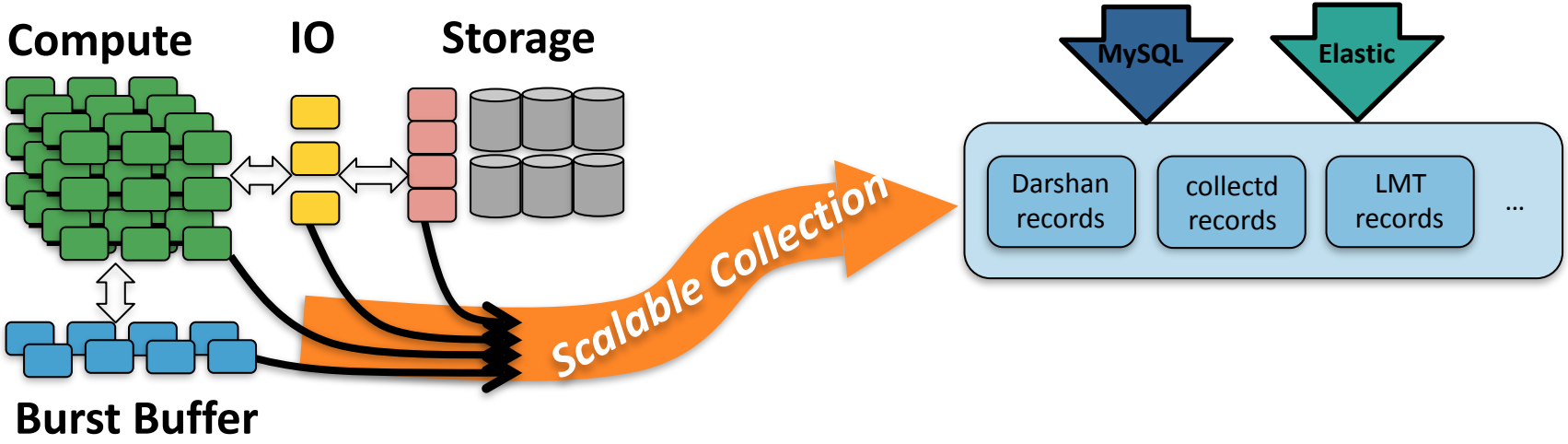


TOKIO: TOtal Knowledge of I/O



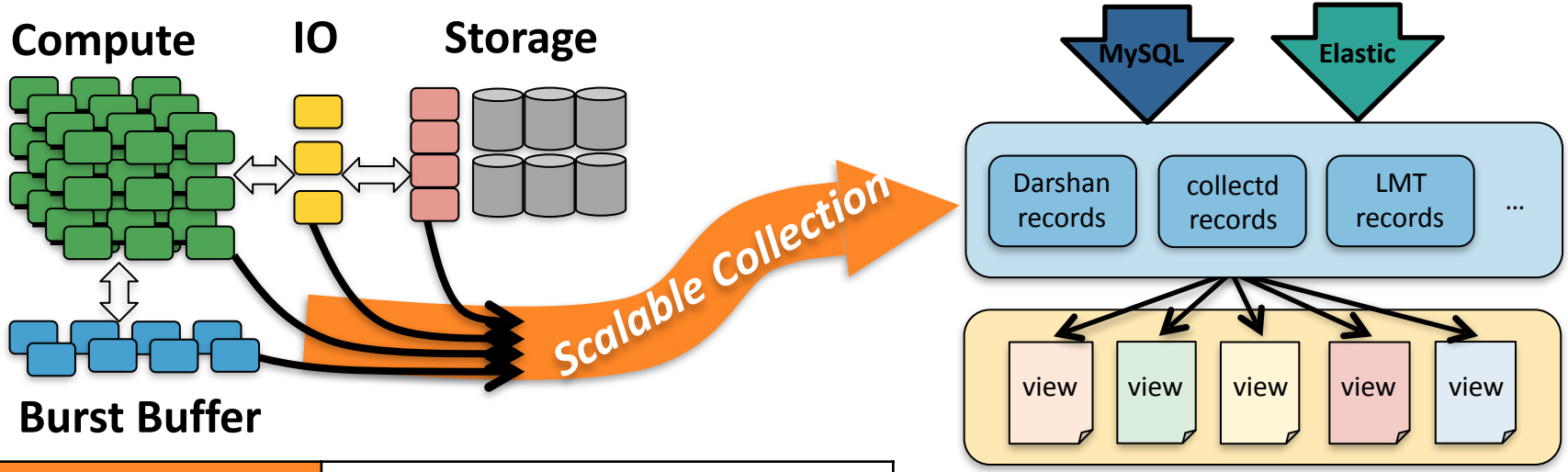
Scalable Collection	RabbitMQ

TOKIO: TOtal Knowledge of I/O



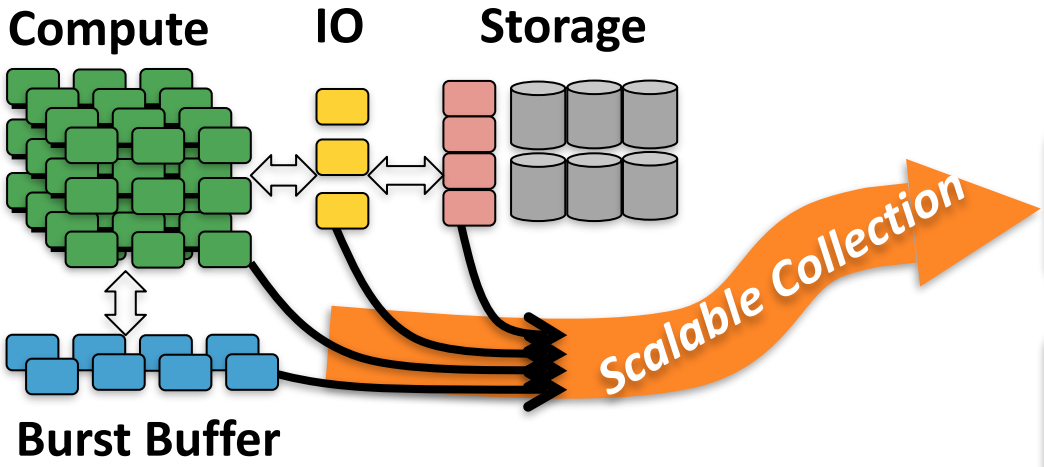
Scalable Collection	RabbitMQ
Record Formats	Darshan logs, HDF5

TOKIO: TOtal Knowledge of I/O

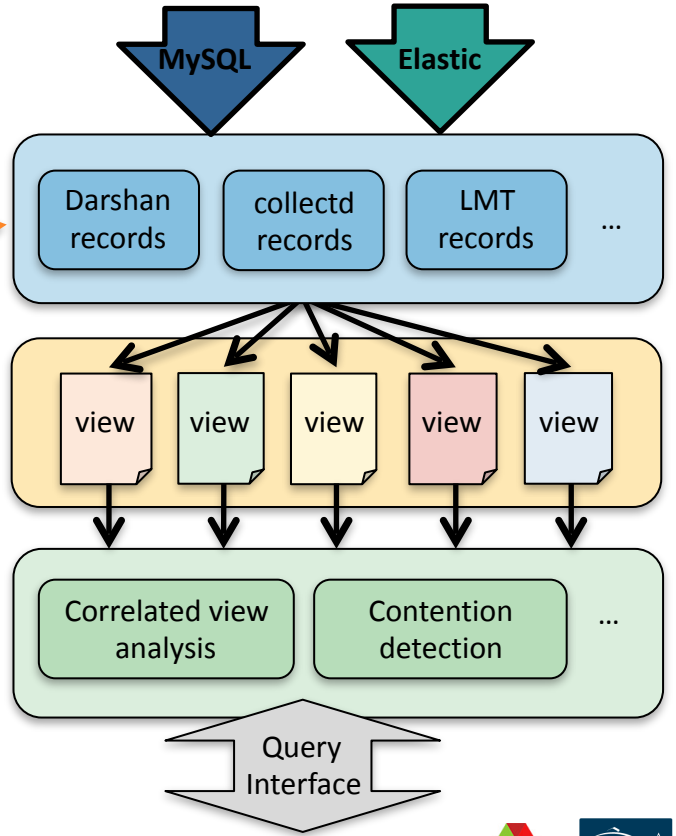


Scalable Collection	RabbitMQ
Record Formats	Darshan logs, HDF5
Views	Index on time, topology, job id

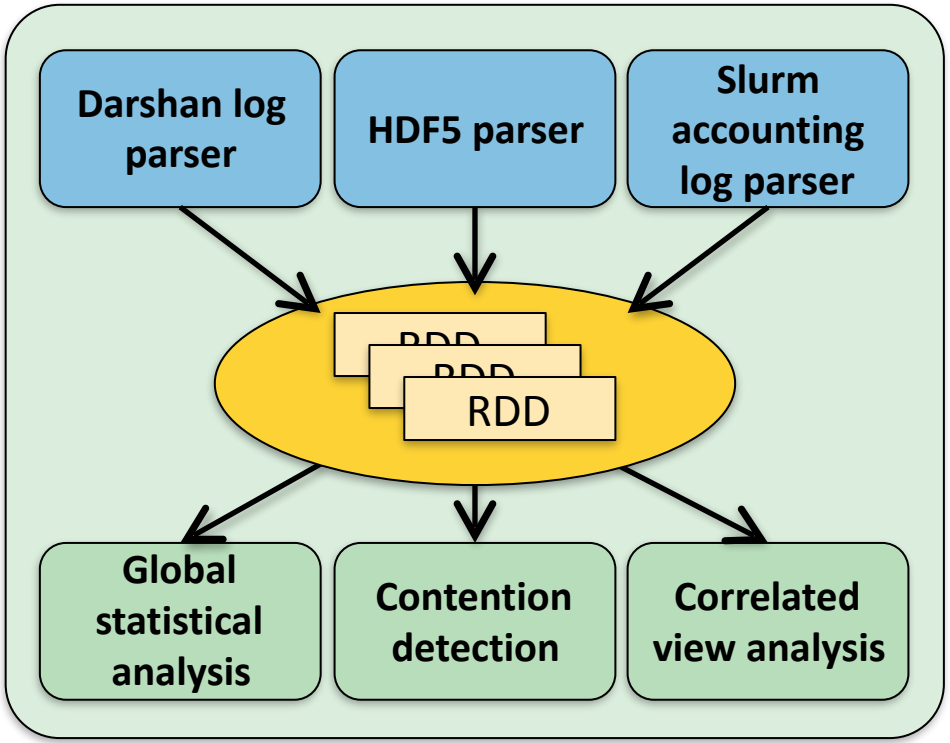
TOKIO: TOtal Knowledge of I/O

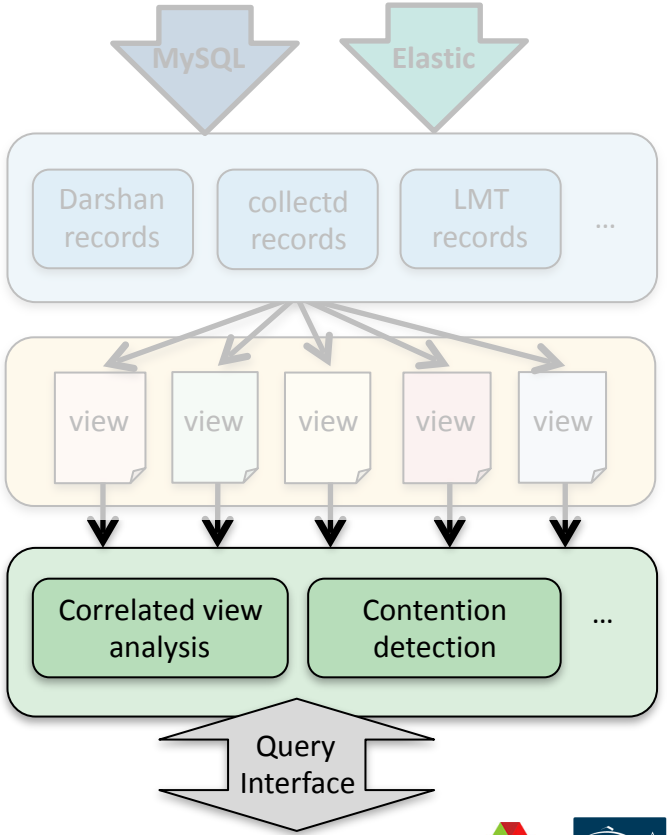


Scalable Collection	RabbitMQ
Record Formats	Darshan logs, HDF5
Views	Index on time, topology, job id
Analysis Modules	Apache Spark

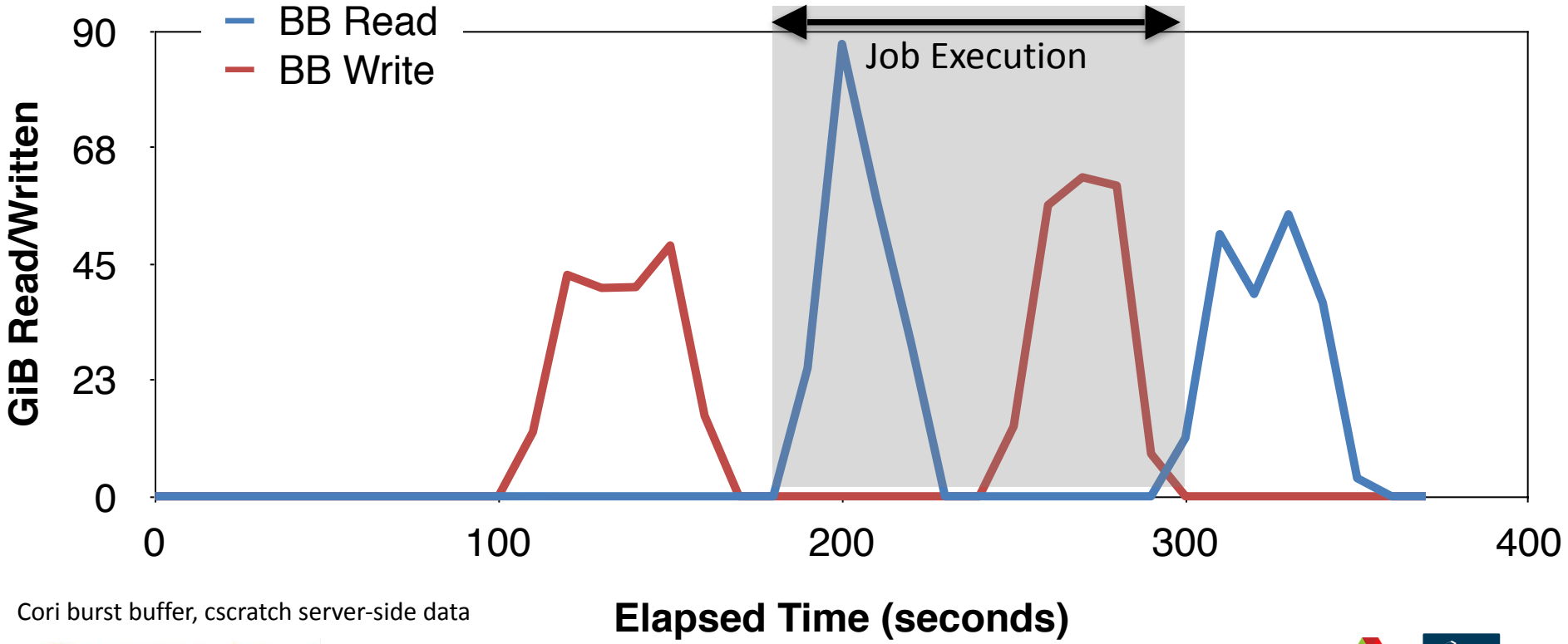


TOKIO Analytics with Spark



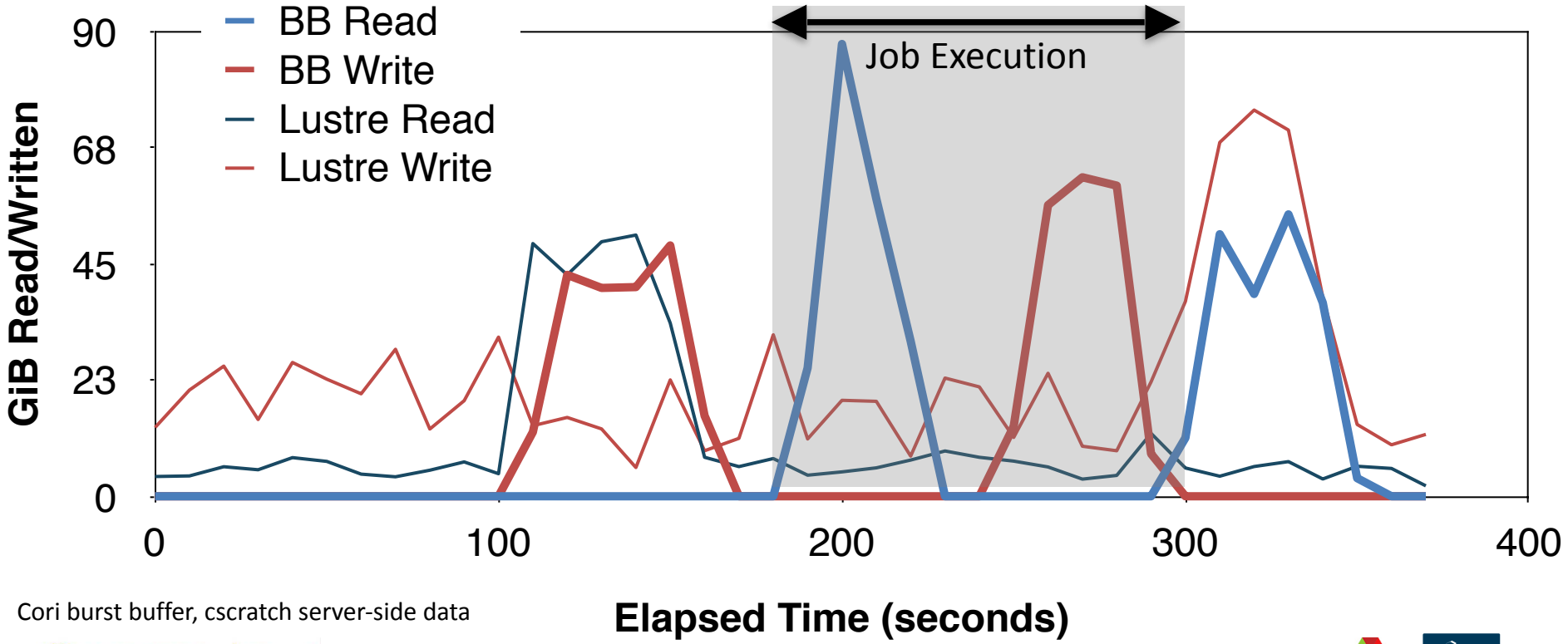


Correlating Lustre & DataWarp Server Metrics



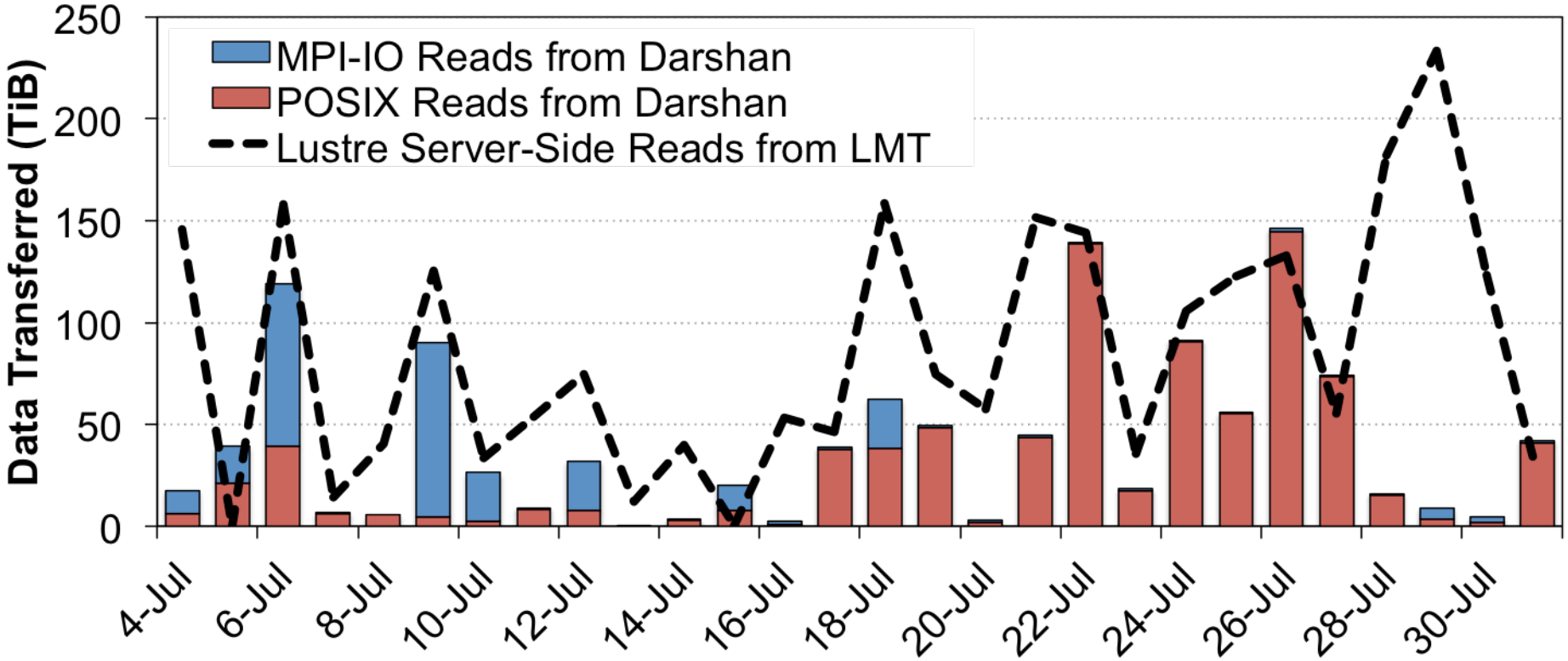
Cori burst buffer, cscratch server-side data

Correlating Lustre & DataWarp Server Metrics



Cori burst buffer, cscratch server-side data

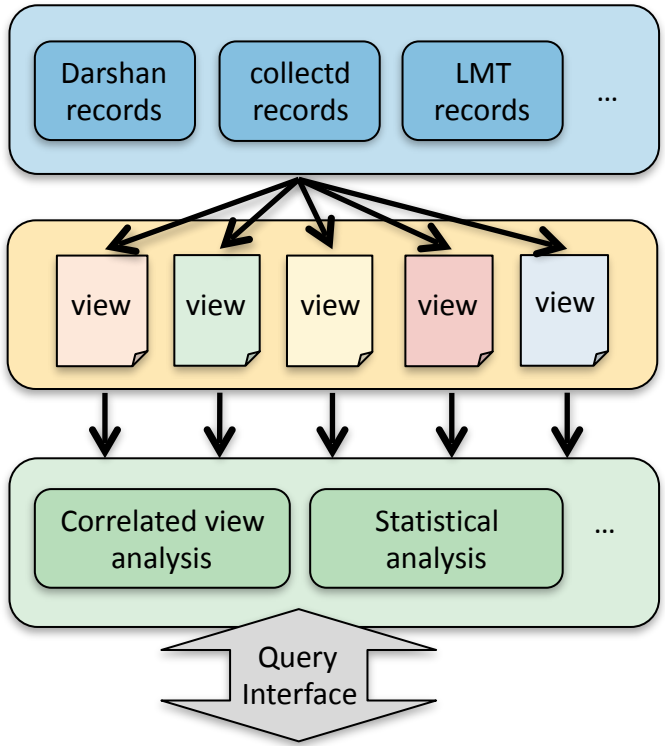
Correlating Darshan & Lustre Server Metrics



Edison Darshan logs, Edison scratch1 server-side data

TOKIO Framework: Ongoing Work

- **Records** – creating standardized HDF5 schema (TOKIOFile™) common to Lustre, GPFS, DataWarp, …
- **Views** – what are the most useful views to maintain?
- **Query interface**
  - REST access to records
  - REST access to Spark analyses
- **Blueprints, libraries, documentation, and data to be publicly available**



Ongoing work funded by the DOE ASCR SSIO program in collaboration with

Shane Snyder
Matthieu Dorier
Philip Carns
Robert Ross
Jialin Liu
Wucherl (William) Yoo
Suren Byna
Nicholas J. Wright