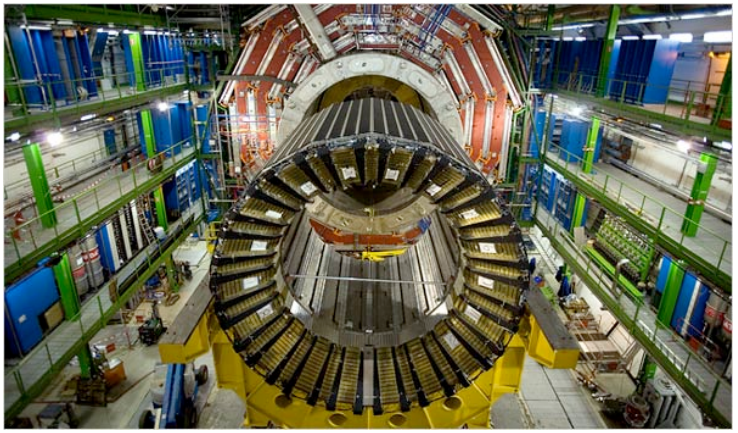
Exploiting Latent I/O Asynchrony in Petascale Science Applications

Patrick Widener, Mary Payne, **Patrick Bridges**
University of New Mexico
Matthew Wolf, Hasan Abbasi, Scott McManus, Karsten Schwan
Georgia Institute of Technology

The research described in this presentation was supported by the National Science Foundation's HECURA program, the Department of Energy's Office of Science, and the U.S. Defense Threat Reduction Agency

THE UNIVERSITY of NEW MEXICO

Data intensities increasing everywhere

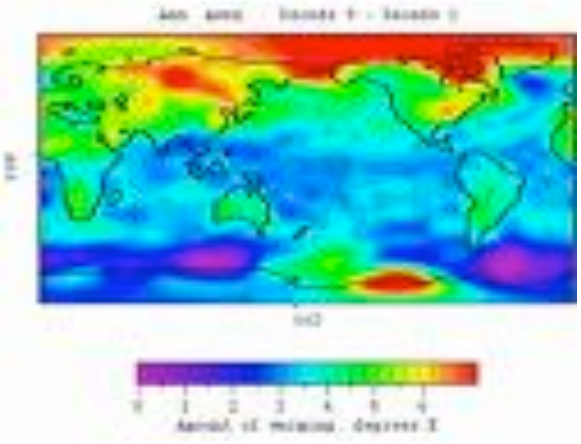


Large Hadron Collider
2 PB/sec

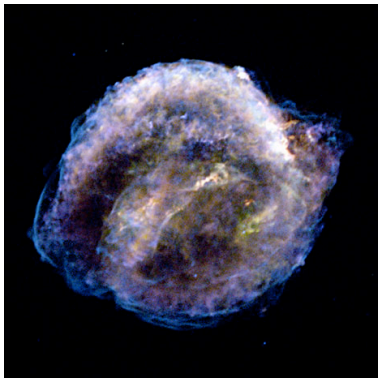
NG power grids
45 TB/day

<u>Storage</u> is challenging, let alone <u>analysis</u>: *write-once, read-never*

Data extract -> store -> analyze/visualize *will not scale*



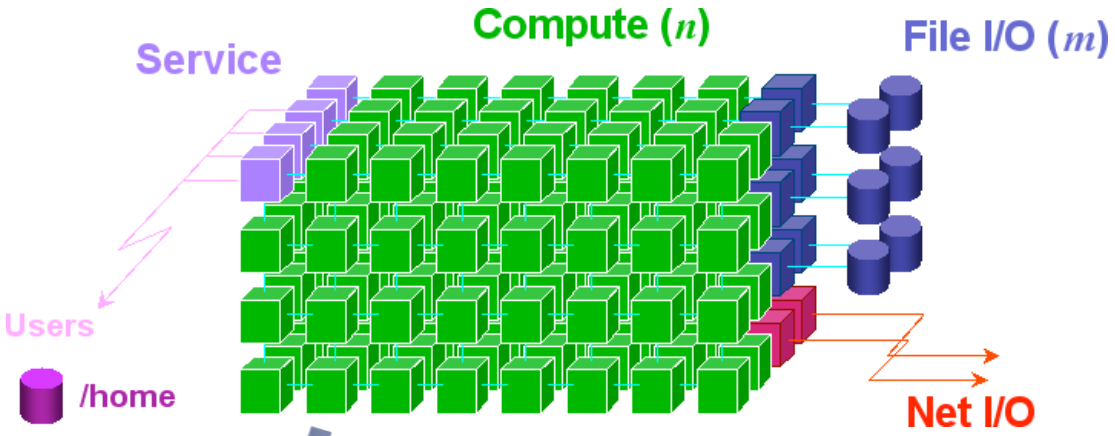
Climate modeling
8 PB/run



ORNL Chimera
35K cores, 550 KB/core/sec => ~18 GB/sec

THE UNIVERSITY of NEW MEXICO

ORNL GTC fusion simulation: 60 TB/run



Gyrokinetic Toroidal Code

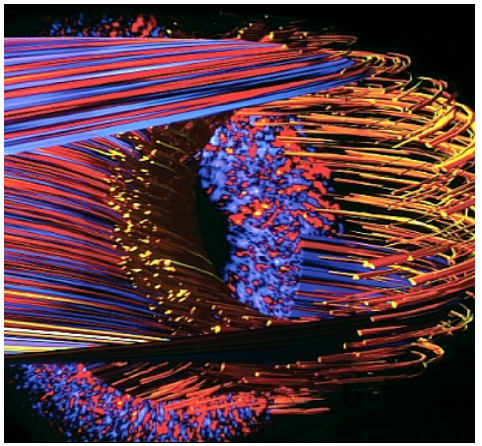
- > 10000 nodes ORNL Cray XT4
- 1024:1 compute / I/O ratio
- Limited I/O node disk BW
- Scarce memory, CPU on compute nodes

Checkpoint / Restart

- Periodic export of all particles (potentially $>10^9$)
- 10% of node memory (200MB/core)
- ~8TB/write on 40K core XT4

Lustre PFS

after run completed



Analysis

- Reorganization, cleaning
- Filtering, extraction
- Monitoring, playback

THE UNIVERSITY of NEW MEXICO

I/O Demands are limiting scientific applications on these systems

Problem: In-band data filtering, transformation, and analysis slows core scientific computation with ancillary tasks

- Thin pipe to I/O subsystem (I/O network, disk spindles) r
- I/O generally synchronous because compute node memory storing the I/O data is scarse
- Metadata updates are frequently slow and often unnecessary
- Lack of systems to enable application scientists to move tasks out of band

THE UNIVERSITY of NEW MEXICO

Decoupled data annotation & processing

Contribution: I/O techniques to decouple filtering, transformation, and analysis from compute nodes

- ***IOgraphs*** decouple data manipulations in space from applications
- ***Metabots*** decouple data manipulations in time *and* space

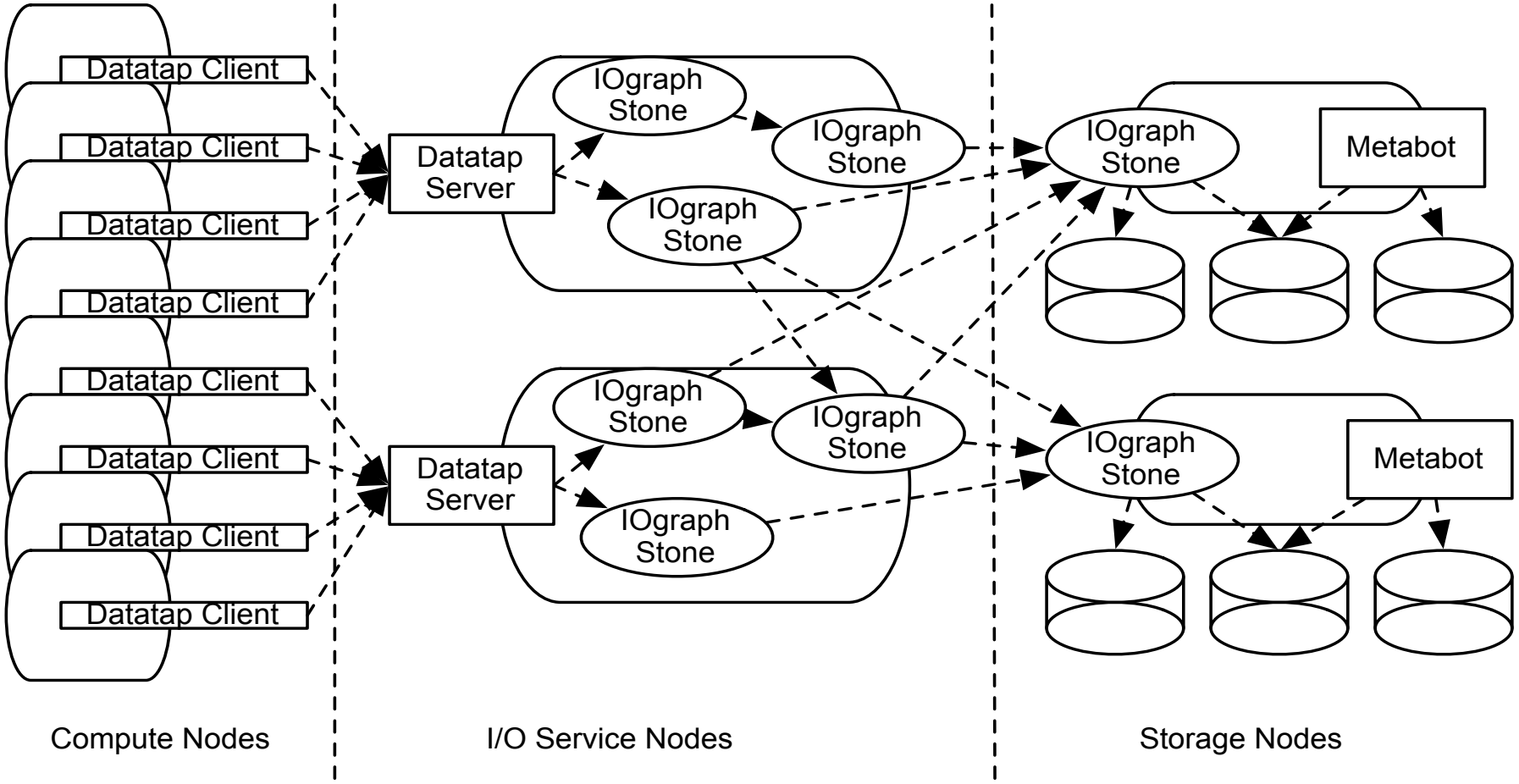
Enabling Technologies:

- ***DataTaps*** export data and "just enough" metadata using a smart, context-aware RDMA transfer
- ***Lightweight File System*** (LWFS) provides minimum filesystem semantics

Using these tools to decouple ancillary operations can improve application I/O throughput, while giving end-users better abstractions to work with

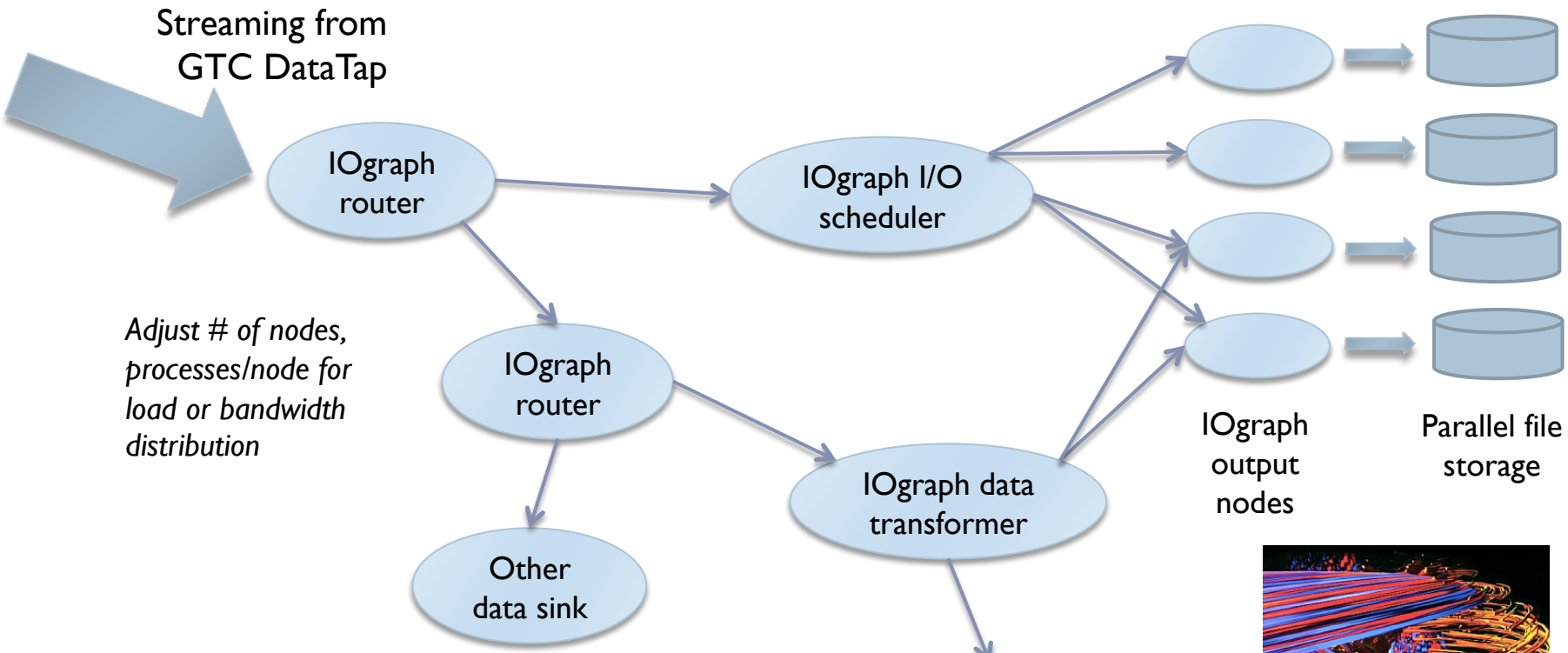
THE UNIVERSITY of NEW MEXICO

Software architecture for "in-transit" data annotation and processing



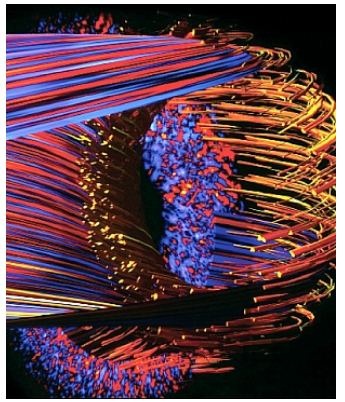
THE UNIVERSITY of NEW MEXICO

IOgraphs decouple operations in space



Act on data in transit

- Dynamic overlay mapped to cluster, non-cluster nodes
- Streaming model, structured data
- Dynamically generated code, shared objects implement operations

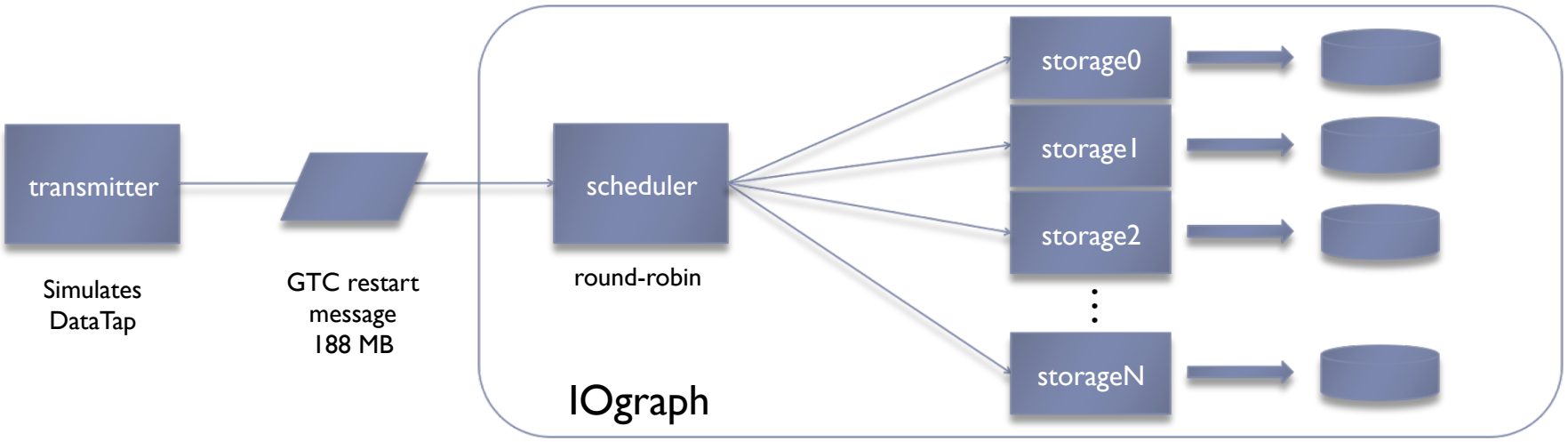


Stream visualization

THE UNIVERSITY of NEW MEXICO

What should IOgraphs look like?

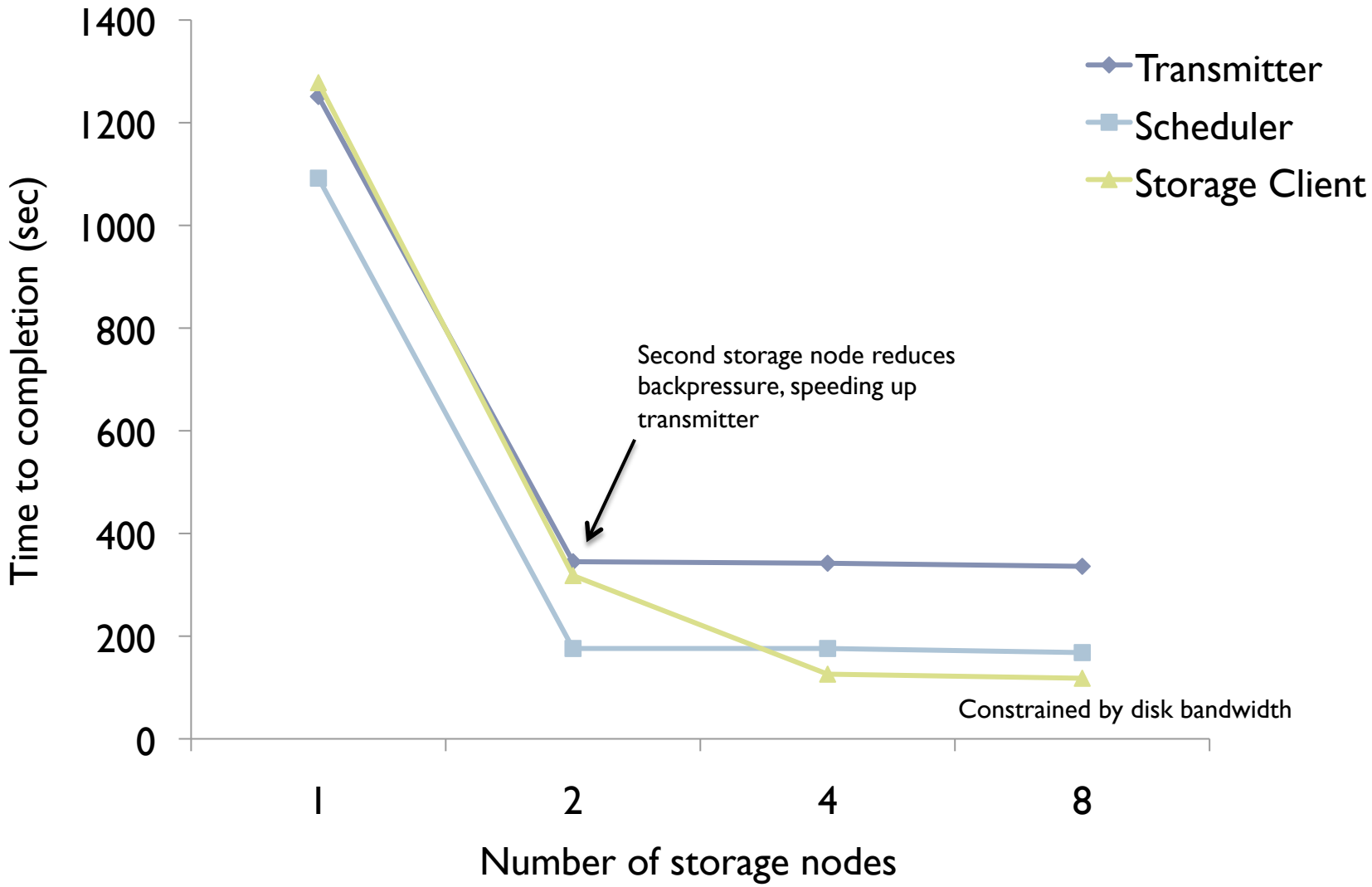
- For buffering and distribution of I/O: # of nodes, # of processes/node?



- Modeling construction of GTC restart file
- Transmitter sends 200 messages
- Scheduler round-robins messages to storage nodes, which write to disk

THE UNIVERSITY of NEW MEXICO

Adding nodes to IOgraph shortens I/O phase



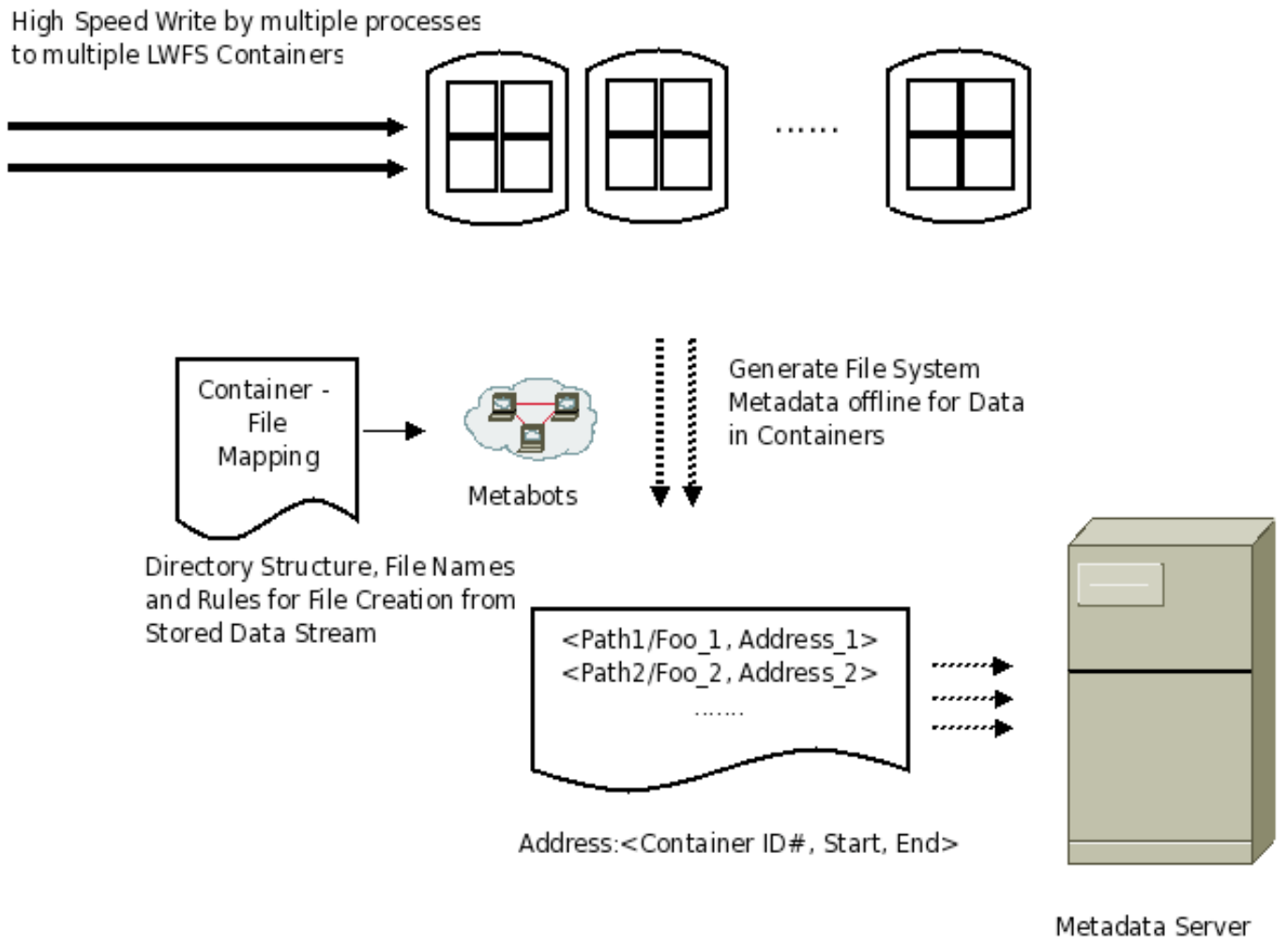
THE UNIVERSITY of NEW MEXICO

Metabots decouple operations in time

- Some operations can or must be delayed
  - Data formatting in long-running MPP codes
  - Some data products may not be needed
  - Service node numbers may be limited or overcommitted
- Small, modular programs; specification-based
  - Well-defined input, output, transformation
  - Data consistency/availability, co-scheduling information
- Ideal for just-in-time, on-demand conversions or metadata fixups
- Use same metadata, transport infrastructure as IOgraphs

THE UNIVERSITY of NEW MEXICO

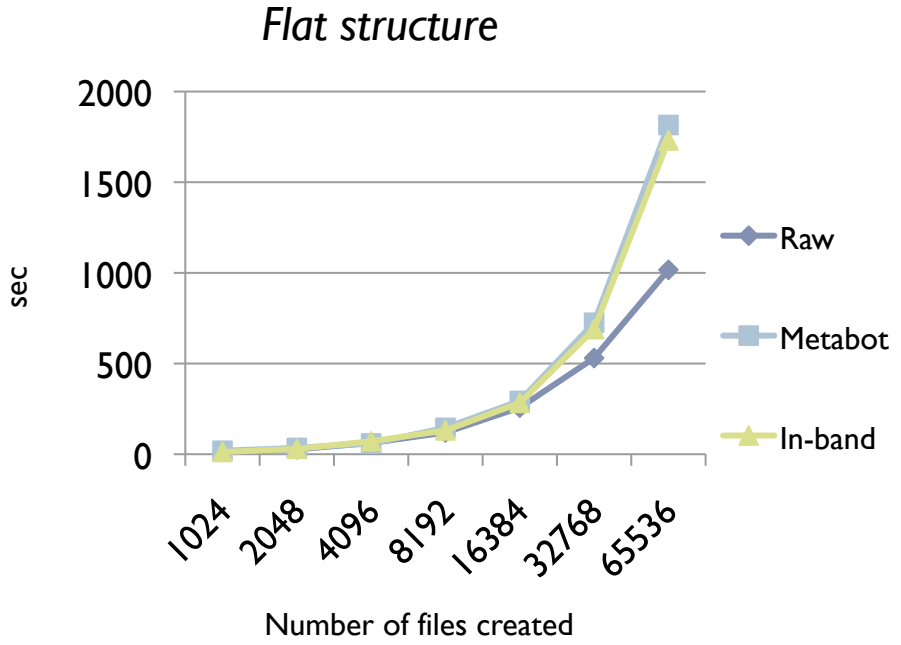
Deferring directory metadata creation

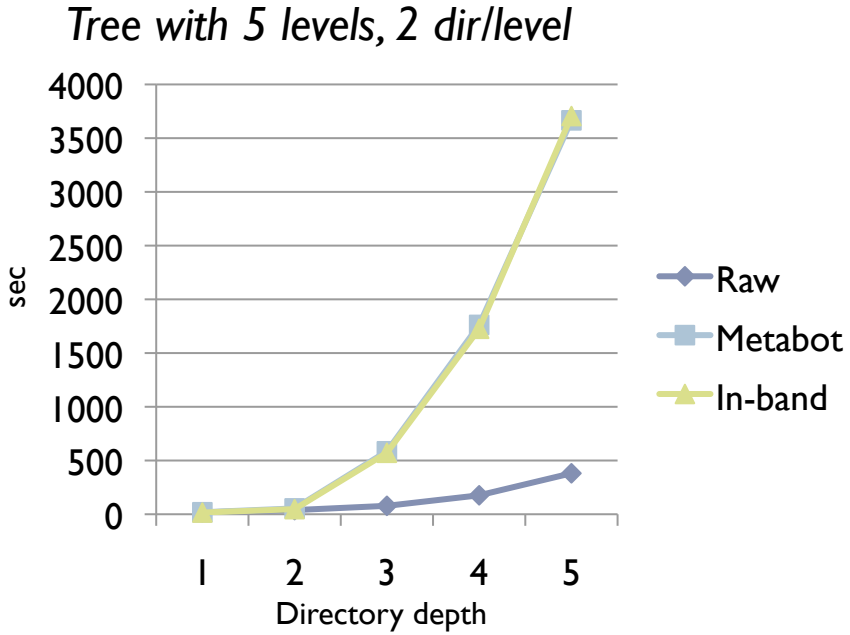


THE UNIVERSITY of NEW MEXICO

Lazy metadata construction reduces wall-clock time

- Create structure without directory information (LANL FDTREE)
- Fix up later (add to LWFS name service) with metabot



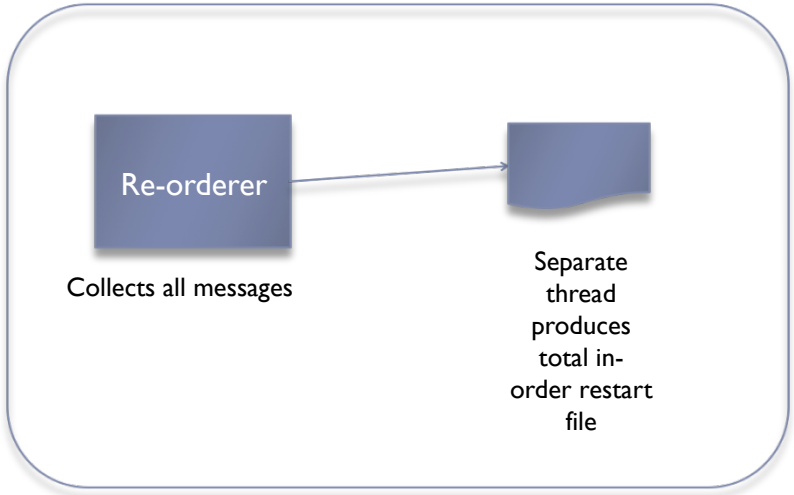


- In-band is 70% slower on flat structure
- In-band is > 9X slower on tree structure
- Metabot reconstruction time similar to in-band time, ***but decoupled***

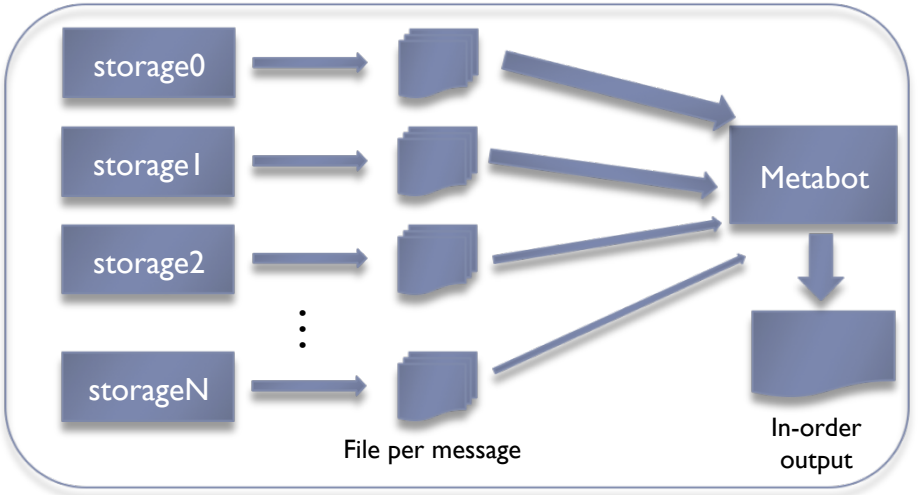
THE UNIVERSITY of NEW MEXICO

Combining IOgraphs and metabots reduces overall execution time

- Create a fully-sorted restart file from collection of messages?
- Single sorter vs. write-now, merge-later



In-band with IOgraph



Metabot

	In-band Processing	Metabot Processing	Total
Single In-series writer/sorter	2113.16	--	2113.16
2 storage nodes + metabot	250.91	526.71	777.62
4 storage nodes + metabot	216.52	526.71	743.23

THE UNIVERSITY of NEW MEXICO

Comparison to other work

- High-performance parallel file systems
  - Many choices: NASD, Panasas, PVFS, Lustre, GPFS
  - Separation of data from metadata supports our approach
  - Manipulating data en route to/from storage
  - Availability of metadata enables better scheduling, staging, buffering decisions
- DataCutter and related tools
  - Similar goals (e.g. customize end-user visualizations)
  - Richer descriptions for filter and transformation, asynchrony
- Out-of-band techniques are similar to workflow systems
  - Kepler, Pegasus, Condor/G, IRODS, others
  - Specifications like Data Grid Language
  - We focus on fine-grain scheduling, tightly-coupled systems, in-band / out-of-band data manipulation
  - Can metabots be workflow actors?

THE UNIVERSITY of NEW MEXICO

These techniques provide traction on data-intensive applications

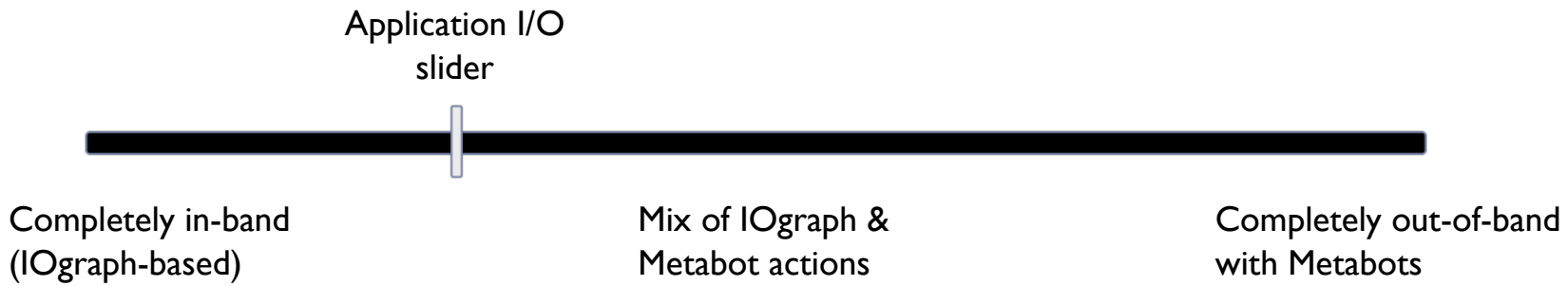
- IOgraphs and metabots provide several benefits
  - Shorten application I/O phases
  - Make analysis easier by making customization easier
  - Reduce net storage amounts
  - Generate custom metadata
  - Accommodate anonymous downstream consumers

Using these tools to decouple ancillary operations can improve application I/O throughput, while giving end-users better abstractions to work with

THE UNIVERSITY of NEW MEXICO

Future Work: Dynamic decoupling

- Run-time scheduling decisions about whether to implement operations in IOgraph or metabots
- Longer-range goal is to incorporate feedback
  - CPU / node availability
  - Network bandwidth
  - Data consistency / availability
  - Anonymous / on-demand consumers



THE UNIVERSITY of NEW MEXICO

Acknowledgements

Greg Eisenhauer, Ada Gavrilovska (Georgia Tech)

Barney Maccabe, Scott Klasky (Oak Ridge National Laboratory)

Ron Oldfield (Sandia National Laboratories)