I/O Graphs: Simultaneous Data Transformation and Movement in High Performance Computing

Center for Experimental Research in Computer Systems

Scott McManus, Chetna Kaur, Swaroop Butala, Fang Zheng, Jay Lofstead, Hasan Abbasi, Matt Wolf, Karsten Schwan

{smcmanus, chetnak, sbutala3, fzheng, lofstead, habbasi, mwolf, schwan} @cc.gatech.edu

College of Computing

Georgia Institute of Technology

Mary Payne, Patrick Widener, University of New Mexico

{mpayne, pmw} @cs.unm.edu

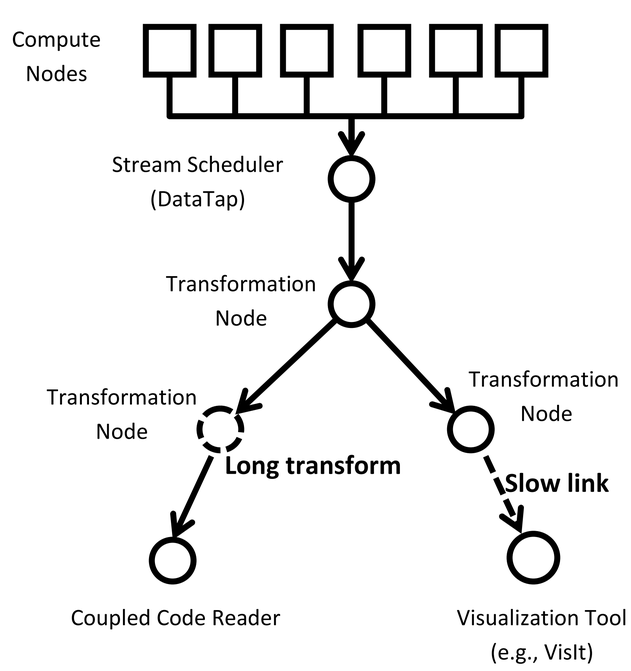
Targeted Applications

- **GTC – Plasma Physics for Tokamak Reactors**
  - Up to 3.6 TB/hour may be transferred.
  - Results subdivided into a mesh for energy visualization
- **XGC0 & XGC1 – Coupled Plasma Physics Codes (Future)**
  - The output from XGC0 must be given to XGC1.
  - Code coupling requires transformation on XGC0's data before it can be used.
- **Chimera – Astrophysics**
  - Restarts (checkpoints) must be saved in multiple formats for later use.
- **Map/Reduce**
  - Potential to change how reduction is performed.

Project Goals

- Transport data efficiently from scientific codes on compute nodes to end users, including visualization tools and other coupled codes.
- Overcome bandwidth limitations in data movement both within and between networks.
- Perform runtime-modifiable data transformations.
- Autonomically migrate operations depending on loads.

Motivational Illustration



Local Decisions - Storage

- I/O Graphs locally decide if messages are arriving faster than they can be handled.
- Data streams in scientific codes are largely periodic, so one possibility is to simply buffer to memory and then to disk.
- One option is to resume the data stream later – this may be preferential for visualization tools with bandwidth constraints.
- Another option is to let a Metabot read from disk later and perform the remaining operations. Operation metadata must be stored with the stored data stream. A reference to the stored data is sent so that each end user can be alerted to read the data.

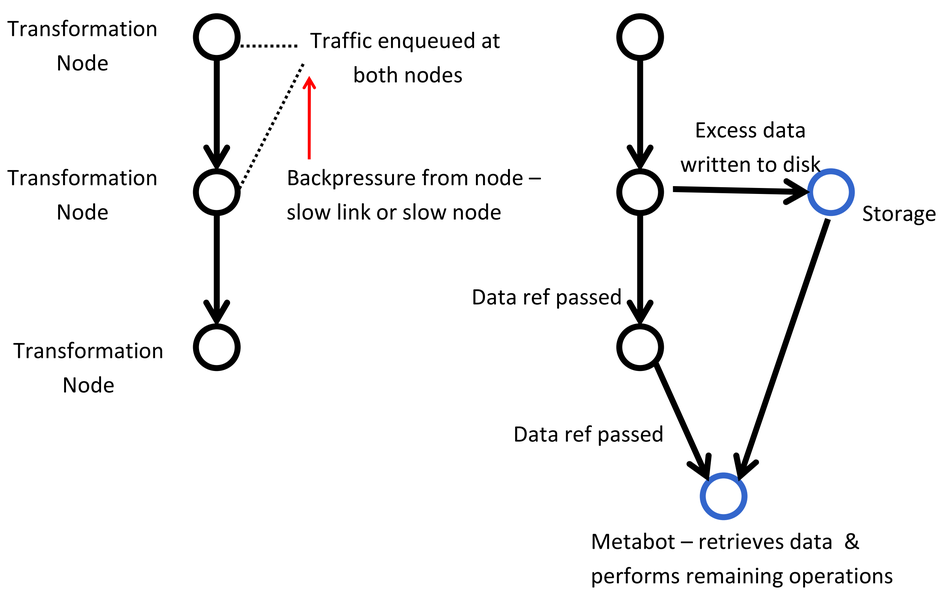
Global Decisions – Control Plane

- Ultimately, we want to reduce dependence on slow secondary storage.
- Long transformations or slow bandwidth often emanates at the source of the data, which makes the local decisions a potentially bad heuristic.
- A control plane for the I/O graph is introduced that may move part of the transformation or duplicate it on more nodes.
- Availability of nodes must be a global decision, possibly between codes.

Operation Metadata & Metabots

- Operations that have been paused and written to disk must include metadata for the functions that have been performed.
- Metabots can operate on this metadata and perform remaining operations on stored data.
- Passing references to the stored data lets Metabots operate independently of graph transformation decisions.

Local Decisions - Example



Global Decisions - Example