**Introduction and Motivation**

- Tremendous amounts of data to process in today’s applications
- Many users have in-house computing resources
  - e.g., local clusters, storage networks
- But the cloud can be used in conjunction to help:
  - Store data remotely
  - Large-scale computation

**Cloud Bursting Challenges**

- Workload may demand more resources over time
- How best to manage a cooperation of cloud and local resources?
  - Data distribution
  - Job scheduling

**Our Processing Framework vs. Map-Reduce**

- We focus on a variant of map-reduce class applications
- **Reduction Object**: Data structure which holds the aggregated result from the reduction phases
- **Local Reduction**: The local reduction function specifies how, after processing one data element, a reduction object is updated.
- **Global Reduction**: The multiple reduction objects are combined to form the final results.

**Cloud Bursting Processing System**

- Master Node at each cluster request a bundle of jobs from the Head Node and assigns each job to the slaves
- Slave Nodes perform the local reduction on the assigned data chunk.
- The assigned data may be from a different cluster
- After all data has been processed, the Head Node invokes the global reduction

**Expiremental Setup**

- **Compute Environment**:
  - **Ohio State cluster**
    - Compute Nodes: Intel Xeon (8 cores) and 6 GB RAM
    - Interconnect: Compute nodes are connected via Infiniband
    - Storage: Dedicated 4TB storage node (SATA-SCSI)
  - **Amazon Web Services cloud**
    - Compute Nodes: m1.large instance (2VC’s, each VC contains 6 elastic compute units = 1.7GHz) and 7.8 GB RAM
    - Interconnect: (High AWS I/O) ??
    - Storage: S3 key-value store

- **Data Intensive Applications and Characteristics**:
  - KNN (low comp, high I/O, small reduction obj)
  - K-Means (heavy comp, low I/O, small reduction obj)
  - PageRank (low comp, high I/O, large reduction obj)
  - 120 GB data for each application

**Experimental Results**

- **Feasibility**
  - Reducing processing overheads
  - Data Retrieval

- **Scalability**
  - Reducing data retrieval overheads
  - Synchronous vs. asynchronous processing

---

*School of Engineering and Computer Science, Washington State University, Vancouver, WA 98686*