Budget-Constrained Bulk Data Transfer via Internet and Shipping Networks

Brian Cho and Indranil Gupta

Department of Computer Science
University of Illinois at Urbana-Champaign
Big Picture: Pandora – People and networks moving data around

• Motivation: Federated cloud with data at diverse geographic locations
  – Research on OpenCirrus

• Data scale of TBs
  – Limited wide area bandwidth is a big bottleneck

• Pandora automates bulk data transfer
  – System autonomically manages transfers across both Internet and physical shipment links
  – Major challenge: create bulk data transfer plans that are fast and cheap
Related Work

• Internet Transfer
  – Grid: [GridFTP]
  – PlanetLab: [CoBlitz 06]

• Disk Shipping Transfer
  – Amazon AWS Import/Export
  – Sneakernets [Jim Gray 03]
  – [PostManet 04], [DOT 06]

• [Cho and Gupta, ICDCS 2010]
  – First ever solution to transfer data cooperatively between multiple sources with internet and shipping edges
    ➢ Better than internet-only and shipping-only strategies
  – Solved for optimal transfer plan that minimizes dollar cost subject to time deadline constraint

This paper: minimum transfer time subject to dollar budget constraint, using [ChoGupta10] as building block.
Option 1: Internet Transfer

Data Source (CMU) → Computation Provider (Amazon)

- 5-20 Mbps
- 1TB: 5-20 days
- $0.10 per GB (AWS)
Option 2: Disk Shipping Transfer

Data Source (CMU)

Data Source (Illinois)

Computation Provider (Amazon)

Disk Interface
40 MB/s

Overnight: $60 per Disk
Two-Day: $30 per Disk
Ground: $10 per Disk

1.5 TB

0.5 TB

No Cost

$0.02 per GB (AWS)
$80 per Disk (AWS)
Budget-Constrained Bulk Data Transfers

• Strict budget constraints
  – e.g., “I have $200 budgeted in my research grant for this transfer of 1 TB to AWS.”

• Internet
  – Cheap (AWS $0.10/GB) but possibly slow
  – e.g., 1 TB dataset transferred from Harvard to Amazon AWS costs $100 but took 3 weeks! [Garfinkel 07]

• Disk Shipment
  – Many per-disk price/time points
  – High per-byte cost if disk is not filled to capacity
    • e.g., 10 GB on 2 TB disk with transfer cost $100: $10.00/GB

• Solve for:
  – Minimum transfer time subject to dollar budget constraint
Budget-Constrained Solution

- Leverage [ChoGupta10]: least cost given a time deadline
- Optimal solution with transfer time $T^*$ is intersection between $B$ and [ChoGupta10] solutions
- Approach: Binary search on [ChoGupta10] solutions
Two-Step Binary Search

- Step 1: Find interval – grow interval by factor of 2
Two-Step Binary Search

- Step 1: Find interval – grow interval by factor of 2
- Step 2: Binary search – shrink interval by factor of 2

Solving for time-deadline solution at each iteration is not cheap
Search is improved by using **bounding functions**

- Bounds the optimal solution to interval $I$

**Bounding function requirements**

- Cost of upper/lower bound must be no less/more than original function
- Computation time for each time deadline should be less than original function
- Interval $I$ should be small
Bounded Binary Search

Search is improved by using **bounding functions**

- Bounds the optimal solution to **interval** $I$

Bounding function requirements

- Cost of upper/lower bound must be no less/more than original function
- Computation time for each time deadline should be less than original function
- Interval $I$ should be small

• Step 1: Find interval for UB
Bounded Binary Search

Step 1: Find interval for UB
Step 2: Binary search for UB

Search is improved by using bounding functions
- Bounds the optimal solution to interval $I$

Bounding function requirements
- Cost of upper/lower bound must be no less/more than original function
- Computation time for each time deadline should be less than original function
- Interval $I$ should be small
Bounded Binary Search

Search is improved by using bounding functions
- Bounds the optimal solution to interval $I$

Bounding function requirements
- Cost of upper/lower bound must be no less/more than original function
- Computation time for each time deadline should be less than original function
- Interval $I$ should be small

- Step 1: Find interval for UB
- Step 2: Binary search for UB
- Step 3: Binary search for LB
Bounded Binary Search

Search is improved by using bounding functions
- Bounds the optimal solution to interval $I$

Bounding function requirements
- Cost of upper/lower bound must be no less/more than original function
- Computation time for each time deadline should be less than original function
- Interval $I$ should be small

- Step 1: Find interval for UB
- Step 2: Binary search for UB
- Step 3: Binary search for LB
- Step 4: Binary search for Original
Bounded Binary Search

Search is improved by using bounding functions
- Bounds the optimal solution to interval $I$

Bounding function requirements
- Cost of upper/lower bound must be no less/more than original function
- Computation time for each time deadline should be less than original function
- Interval $I$ should be small

- Many $UB$ and $LB$ computations
- Less Original function computations
- In worst case:
  - Two-step binary search: $2[\log(T^*)] + 1$
  - Bounded binary search: $[\log(I)] + 1$
[ChoGupta10] uses a time-expanded network (TEN) to solve for least cost given a deadline.

For each binary search iteration, a new TEN is created and a min-cost max-flow solution is computed.

For Bounded Binary Search, we create modified LB/UB time-expanded networks.

Intuitively, incorporate time into graph to create an extended graph representation:
- Make $T =$ deadline copies of each vertex
- Draw edges according to transit time
- Draw holdover edges
Time-expanded Network

- [ChoGupta10] uses a time-expanded network (TEN) to solve for least cost given a deadline
- For each binary search iteration, new TEN created and min-cost max-flow solution computed
- For Bounded Binary Search, we create modified LB/UB time-expanded networks

Intuitively, incorporate time into graph to create an extended graph representation
  - Make $T$=deadline copies of each vertex
  - Draw edges according to transit time
  - Draw holdover edges
Lower Bound (LB)

Time-expanded Network

- Original time-expanded network (TEN): can push 3 units of flow
- LB network: can push up to 4 units of flow
  - Unrealistically optimistic
- Size of network grows with $T/k$, not $T$
- TEN network plan cannot be cheaper than LB
  - $C_{lb}(T) \leq C(T)$
  - Proof in paper
Lower Bound (LB)

Time-expanded Network

• Original time-expanded network (TEN): can push 3 units of flow

• LB network: can push up to 4 units of flow
  – Unrealistically optimistic

• Size of network grows with T/k, not T

• TEN network plan cannot be cheaper than LB
  – $C_{lb}(T) \leq C(T)$
  – Proof in paper
Upper Bound (UB)  
Time-expanded Network

• Original time-expanded network (TEN): can push 3 units of flow
• UB network: can push only 2 units of flow  
  – Lazily pessimistic

• Size of network grows with $T/k$, not $T$
• TEN network plan cannot be worse than UB
  – $C(T) \leq C_{ub}(T)$
  – Proof in paper
Upper Bound (UB) 
Time-expanded Network

- Original time-expanded network (TEN): can push 3 units of flow
- UB network: can push only 2 units of flow
  - Lazily pessimistic
- Size of network grows with $T/k$, not $T$
- TEN network plan cannot be worse than UB
  - $C(T) \leq C_{ub}(T)$
  - Proof in paper
Experimental Evaluation

• Trace-driven
  – Wrote scripts to communicate with FedEx web services: queried package rates and destination time
    • Service at http://hillary.cs.uiuc.edu to do this autonomically
  – Internet BW from PlanetLab measurements
  – 2 TB across 11 source sites: uniform (0.18 TB/source) results shown

• Suitability of UB and LB networks
• Compare binary search strategies
Suitability of UB and LB networks

- Computation time comparatively cheaper
- UB/LB networks reduce interval $I$ of binary search on original network
Bounded Binary Search Strategy

- Bounded Binary Search is better when
  (a) Time within interval reduced
  (b) Time to find interval reduced
Conclusion

• Pandora automates creation of bulk data transfer plans

• Solved problem of finding minimum transfer time subject to dollar budget constraint

• Binary Search on min-cost solutions
  – Two-step/Bounded binary search
  – LB/UB functions using time-expanded networks