Data Distribution and Scheduling for Distributed Analytics Tasks

Stephen Pasteris (UCL), **Shiqiang Wang** (IBM US), **Christian Makaya** (IBM US), **Kevin Chan** (ARL), **Mark Herbster** (UCL)

Time-Critical Analytics Tasks

- Analytics applications are data-driven
- Each analytics task needs to be completed within a specific amount of time

Problem Formulation

System

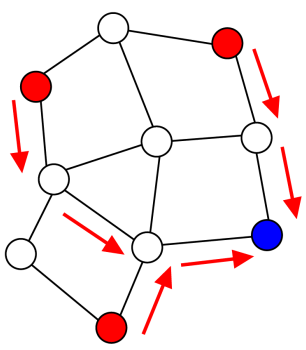
- We have a set of machines V
- Machines can communicate with each other:
  - $B(i,j)$ is the bandwidth of the communication pipeline from i to j
  - There is no propagation latency on the communication pipelines
- S_i is the storage capacity of machine i

Application

- We have a set of applications W
- Each application k must be run on a given machine $a(k)$
- Each application k needs D_k bits of information to run
  - The D_k bits of information need not be stored on machine $a(k)$ but can be sent to $a(k)$ before running the application
- Each application k has a max. task time T_k

Information Streams

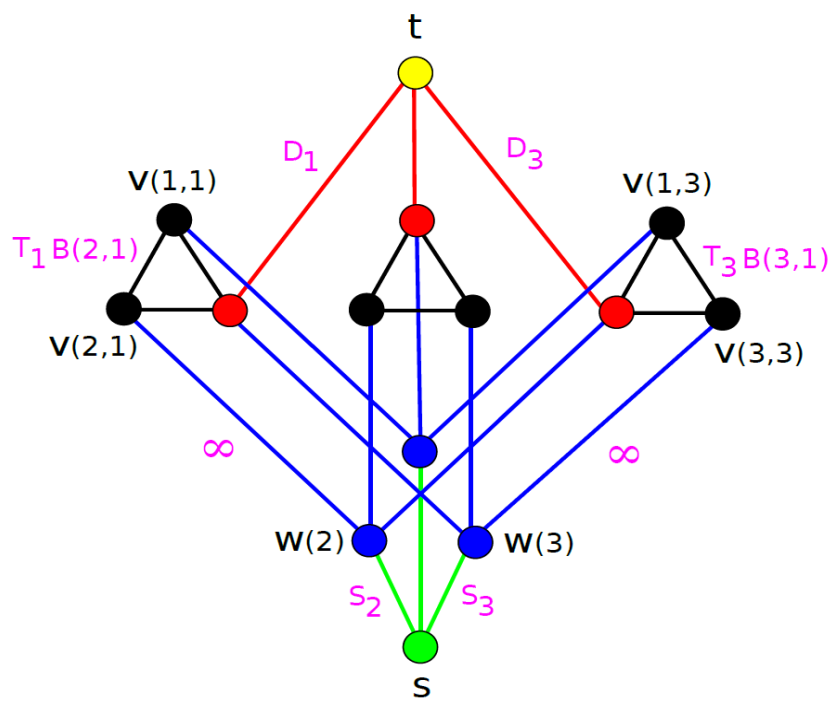
- When we need to run task k all D_k bits of information need to be sent to $a(k)$:
- Let $f(i,j)$ be the rate of information transmitted from machine i to machine j; note that $f(i,j) \leq B(i,j)$
- For a machine i define $\mathrm{l(i)} \coloneqq \sum_{\mathrm{j}\in\mathrm{V}} \mathrm{f(i,j)} - \sum_{\mathrm{j}\in\mathrm{V}} \mathrm{f(j,i)}$
  - $\mathrm{l(a(k))} \leq 0$
  - $\mathrm{l(i)} \geq 0$ for $\mathrm{i} \neq \mathrm{a(k)}$
- Time taken to send all information is $\max_{\mathrm{i}\in\mathrm{V}} d(k,i)/l(i)$; we call this ``**task time**"



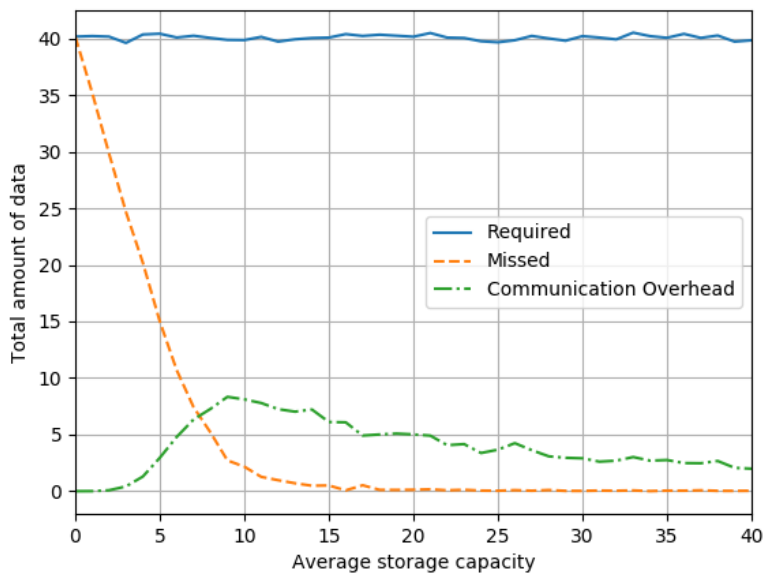
Goal: Find a distribution $d(\cdot,\cdot)$ of the data such that we satisfy the task time requirements of all applications

Algorithm

- Create a source vertex s and a sink vertex t
- For all tasks k and machines i and j create:
  - Vertex $v(i,k)$
  - Vertex $w(i)$
  - Edge from $v(i,k)$ to $v(j,k)$with capacity $T_k B(i,j)$
  - Edge from s to $w(i)$ with capacity S_i
  - Edge from $w(i)$ to $v(i,k)$ with infinite capacity
  - Edge from $v(a(k),k)$ to t with capacity D_k
- Find a maximum flow F from s to t
- There exists a feasible solution if and only if $F(v(a(k),k),t) = D_k$
- We choose $D(i,k)$ equal to $F(w(i),v(i,k))$



Simulation Results



Distributed Analytics and Information Science International Technology Alliance
Annual Fall Meeting, September 2017