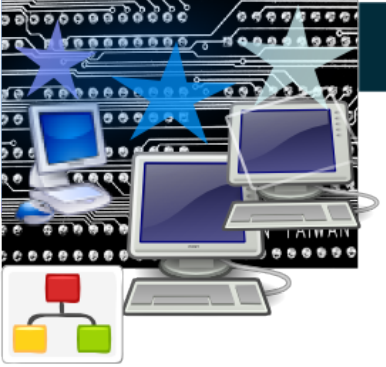


A new heuristic for broadcasting in cluster of clusters – AOC Team –

Christophe Cérin[1], Hazem Fkaier, Luiz-Angelo Steffenel, Mohamed Jemni

[1]Université de Paris XIII, CNRS UMR 7030, France

GPC conference, Hualien, Taiwan

Table of contents

Broadcasting

MPI classroom

4 – Communications collectives : diffusion générale 38

4.3 – Diffusion générale : MPI_BCAST()

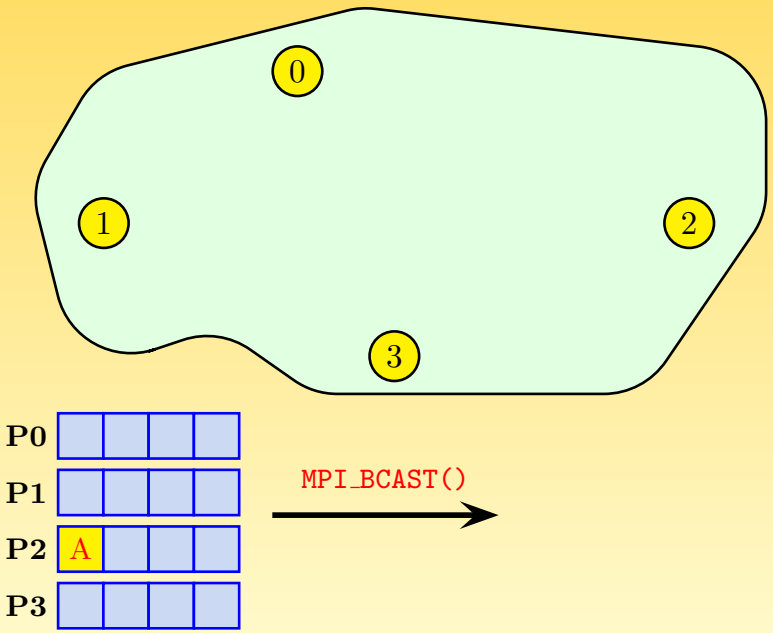


FIGURE 12 – Diffusion générale : MPI_BCAST()

Broadcasting

MPI classroom

4 – Communications collectives : diffusion générale 38-a

4.3 – Diffusion générale : MPI_BCAST()

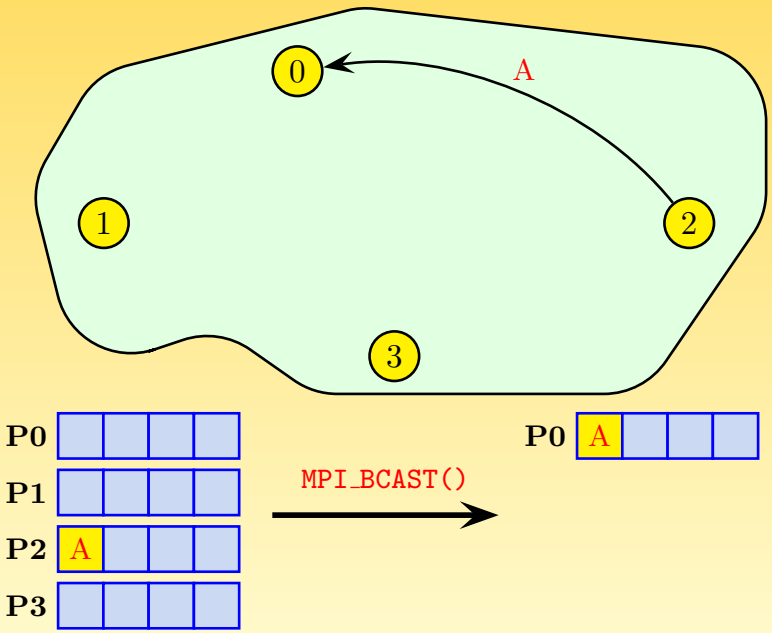


FIGURE 12 – Diffusion générale : MPI_BCAST()

Broadcasting

MPI classroom

4.3 – Diffusion générale : MPI_BCAST()

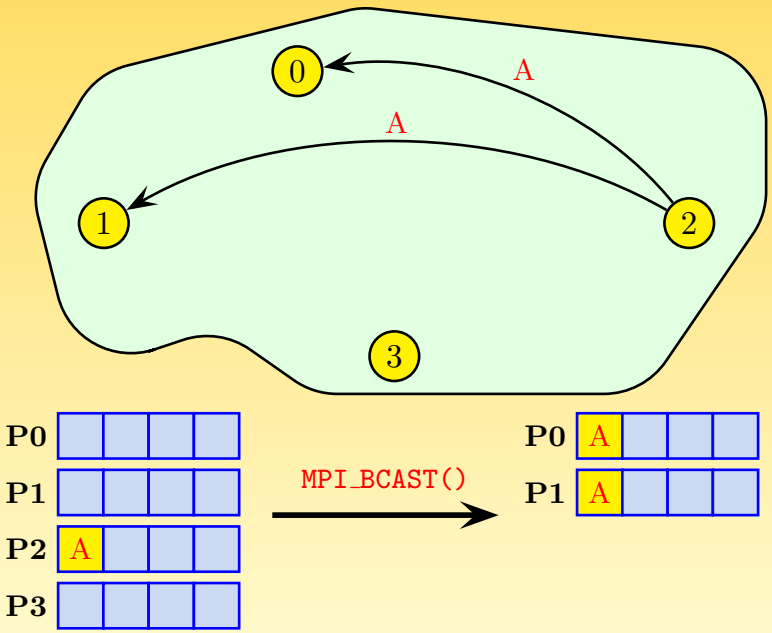


FIGURE 12 – Diffusion générale : MPI_BCAST()

Broadcasting

MPI classroom

4 – Communications collectives : diffusion générale 38-c

4.3 – Diffusion générale : MPI_BCAST()

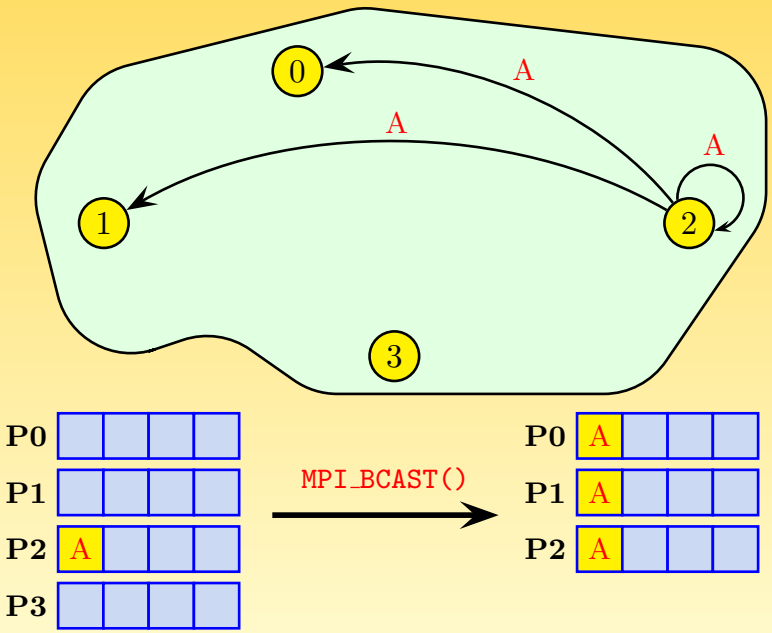


FIGURE 12 – Diffusion générale : MPI_BCAST()

Broadcasting

MPI classroom

4.3 – Diffusion générale : MPI_BCAST()

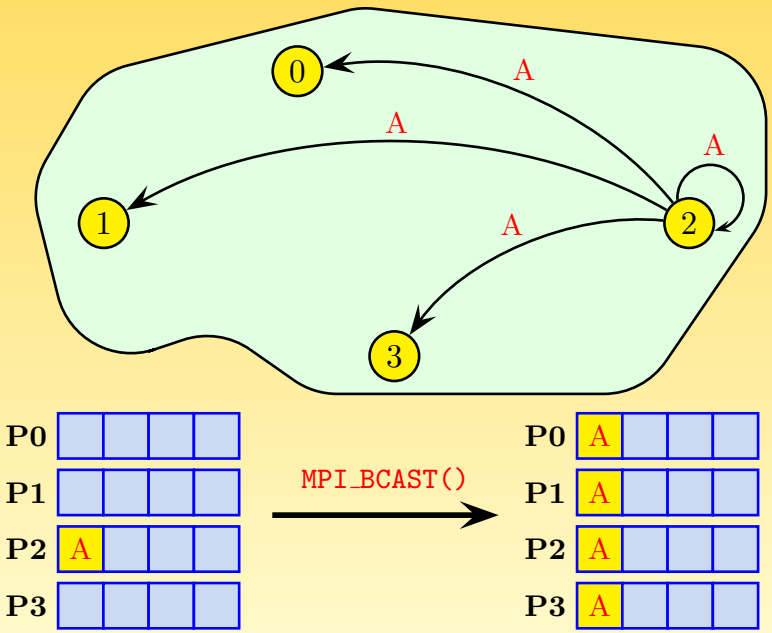
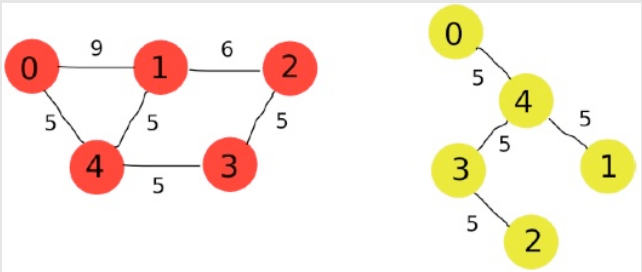


FIGURE 12 – Diffusion générale : MPI_BCAST()

Broadcasting

Tree construction

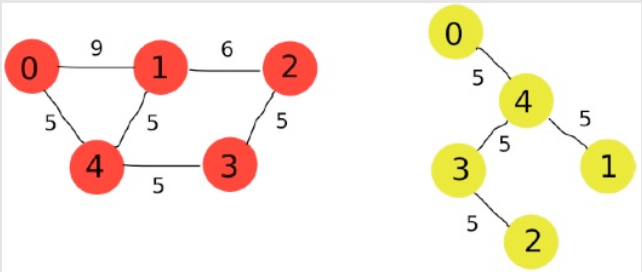


Key Points

- Cost on vertices;
- Multi-port model (?);
- Optimization: redistribution;

Broadcasting

Tree construction



Key Points

- Cost on vertices;
- Multi-port model (?);
- Optimization: redistribution;
- Take into account the physical hierarchy: cluster of cluster $\neq$ flat organization

- Related work: techniques, hypothesis...

Derive complexity results

- Difficult problem: finding the tree that minimize the execution time is NP-complete even for the basic telephone model (problem ND49 in Garey & Johnson)
- Performance metric: maximizing the throughput? (i.e the max streaming rate, once steady state has been reached)

Do not derive math results

- Metric: ensure no more than N messages (simultaneous);
- Emulation / simulation: simple implementation / complex one (a 'theoretical' algorithm may require lot of resources)

Related work: techniques, hypothesis...

Theoretical bounds on the metrics

- Model of the architecture: Internet like (complete connectivity) ; meshes ; hypercubes...
- Model of communication: a processor can be involved simultaneously in several communications (in-out bandwiths are not exceeded)
- Other assumptions: contentions...

Approximate the solution

- Horror: we may be arbitrary far from the optimal solution;
- Heuristics: Fastest Node First...

Related work: Broadcasting on Large Scale Heterogeneous Platforms under the Bounded Multi-port Model

Beaumont & al - Ipdps'2010

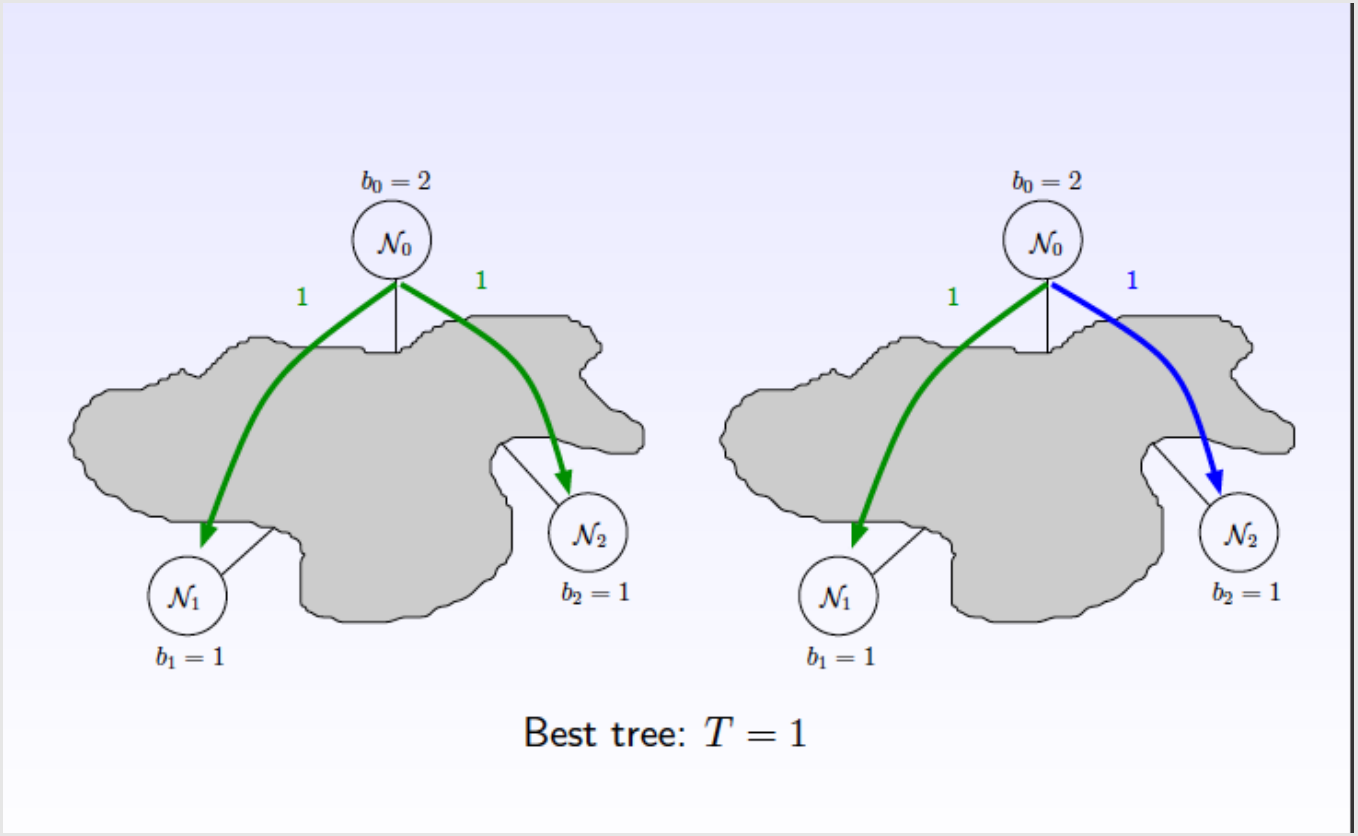
- Overlay: that make the best possible use of the communication capabilities of all nodes;
- Model of communication: all nodes may communicate with the others ; the backbone is large enough $\Rightarrow$ contention 'localized' on the nodes $\Rightarrow$ represent the platform by local properties (in-outgoing bandwidths);

Discussion

- Implementation: multi port model may use TCP QoS mecanisms (bandwidth sharing) $\Rightarrow$ Comet server or Ajax Push Engine are fine for 100K nodes, not 10^{18} nodes! $\Rightarrow$ but the solution is hierarchical!
- Good test: Technological limits of Comet
- Simulation: no experimental results on large platforms

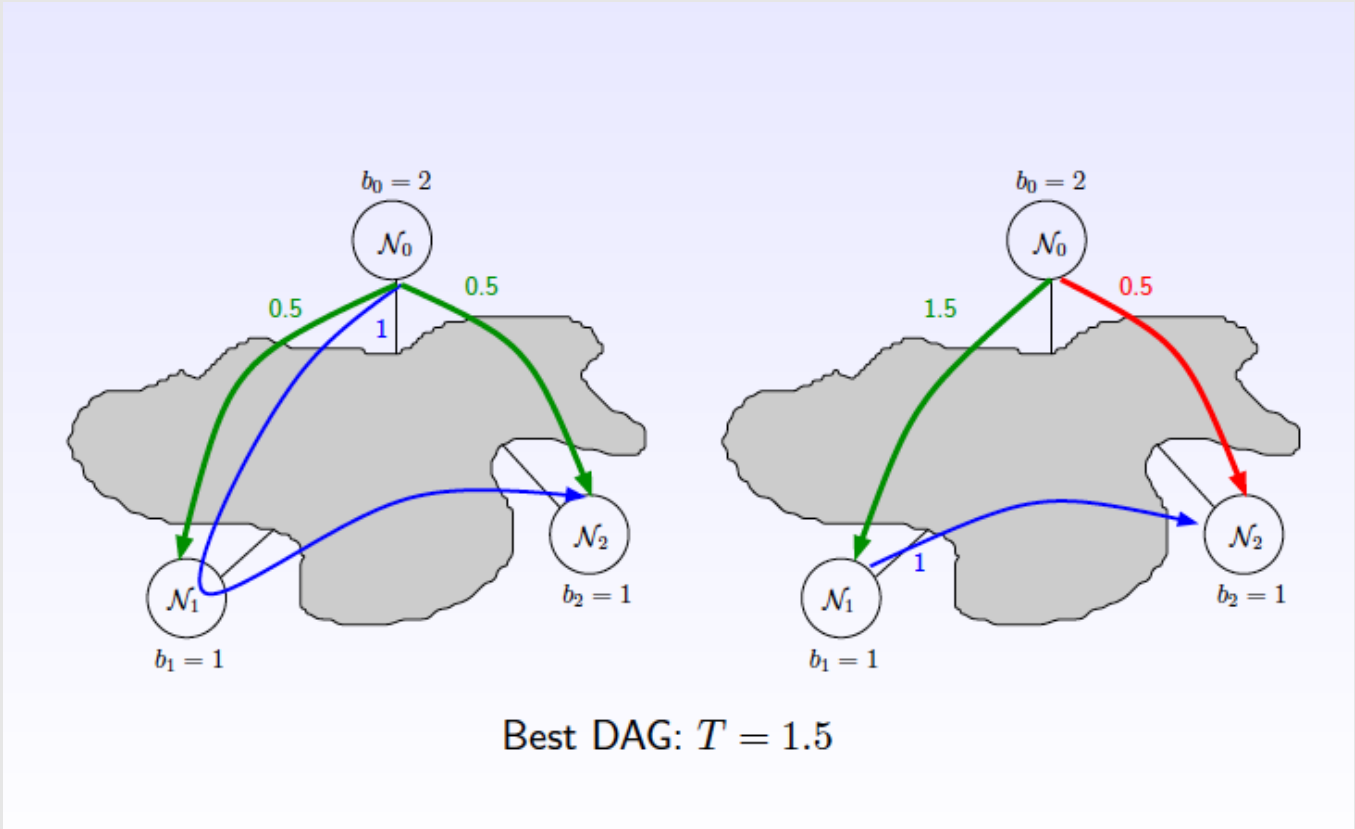
Broadcasting and throughput issues

Olivier Beaumont examples (1/3)



Broadcasting and throughput issues

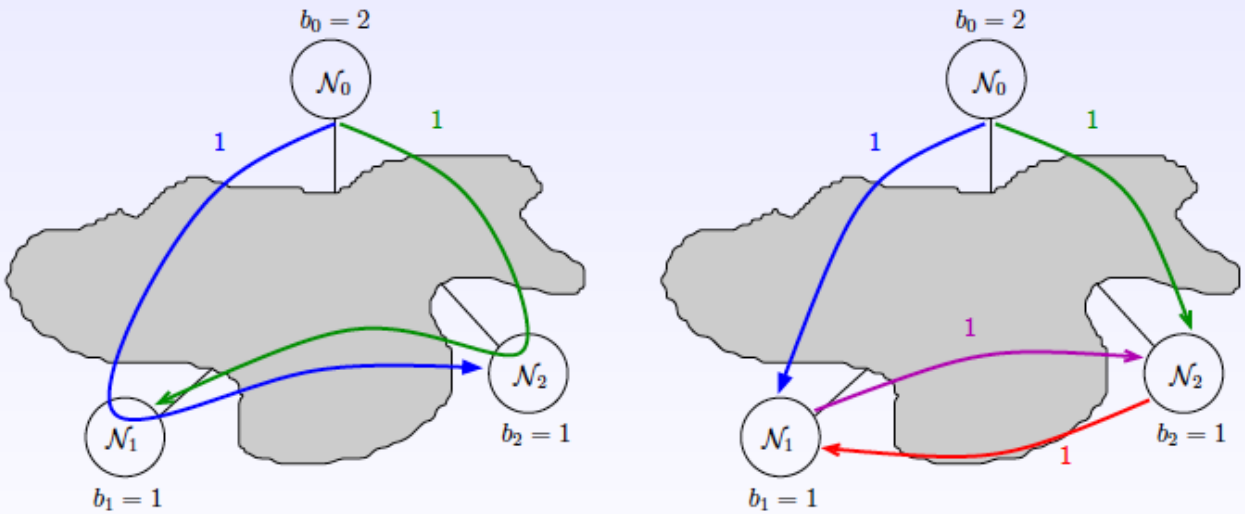
Olivier Beaumont examples (2/3)



Best DAG: $T = 1.5$

Broadcasting and throughput issues

Olivier Beaumont examples (3/3)



Optimal: $T = 2$

Once the overlay is computed, there exists distributed algorithms to perform the broadcast.

Scheduling for atomic broadcast operation in heterogeneous networks with one port model

Robert Hsu & al - The Journal of Supercomputing, Dec 2009

- Results: a) a graph based alg for homogeneous environment; b) a tree based alg. for heterogeneous environments $\Rightarrow$ minimizing makespan
- Explanations: tree based means the avoidance of duplicated messages ; graph based means we do not remove communication links in the original graph;

Discussion

- Implementation: yes (heuristics)
- Simulation: use a random graph generator
- Large scale experiments: no

Scheduling for atomic broadcast operation in heterogeneous networks with one port model

Robert Hsu & al - The Journal of Supercomputing, Dec 2009

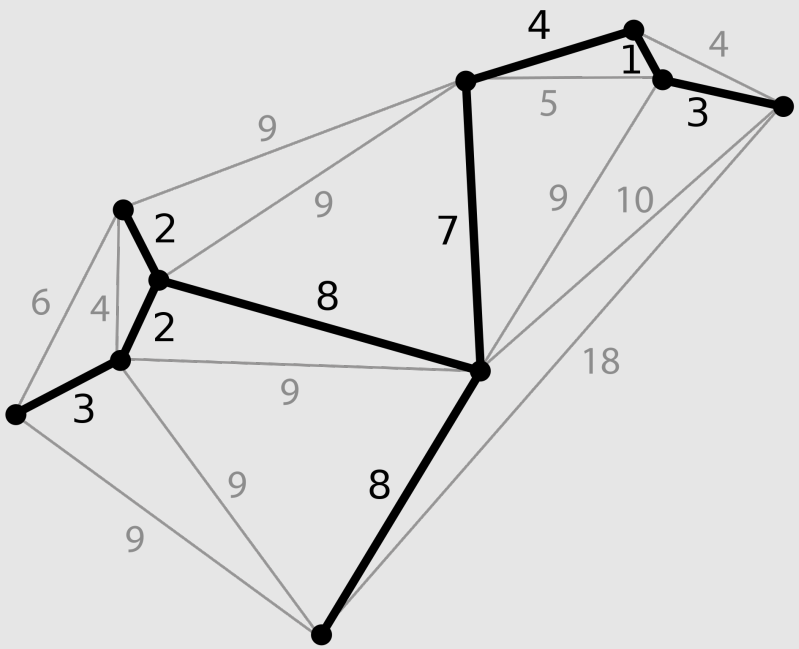
- Graph based: NNF (Nearest Neighbor First) ; MDNF (Maximum Degree Neighbor First)
- Tree based: Tree-based Nearest Neighbor First scheduling (MST + NNF) ; Tree-Based Maximum Degree Neighbor First scheduling (MST + MDNF) ; Tree-Based Maximum Height Subtree First scheduling ...

Key point for the tree based approach

- Given a minimal cost spanning tree that reduced from an original Het. Network.
- MST: does not guaranty the optimality! A guidance for the unicity of paths + redistribution
- Heuristics: all the alg.

Broadcast Tree vs Minimum Spanning Tree

Example



In the optimal broadcast problem the issue is to minimize the time to reach the last node that minimizes the longest path in the tree. In the MST, the issue is to minimize the whole 'weight' of the tree.

Origins of the work (heterogeneous cluster)

References

Basic principle: from set B (nodes without the message) to set A (nodes with the message)
Note: $P_i \in A, P_j \in B$

Early Completion Edge First - ECEF: based on $RT_i + g_{ij}(m) + L_{ij}$. Goal: increase number of nodes in set A as fast as possible.

Early Completion Edge First with look-ahead - ECEF-LA: Bha proposes to estimate the efficiency of each node throughout a function that takes into consideration the speed of **forwarding** the message to another node of set B.

$$F_j = \min_{P_k \in B} (g_{jk}(m) + L_{jk})$$

Grid aware alg. (cluster of clusters)

References

Luiz Angelo: a proxy on each cluster and the local communication load is depicted by

$$L_{kk'} + g_{kk'} = \begin{cases} T_k & \text{if } k' \text{ is associated to } k \\ \infty & \text{if } k' \text{ is not associated to } k \end{cases}$$

BottomUp: starts by contacting the most loaded proxy. The heuristic needs to contact it through the 'shortest path'. BottomUp uses a min-max approach to find the *'shortest path'* to contact the most loaded coordinator.

The new approach (cluster of clusters)

Take into consideration 3 factors

1) We need to increase the size of set A with clusters, in the quickest possible way;
2) Availability of numerous senders give us more chance to perform next communication in a better way, since we have more choices to consider (advantage to communication-effcient clusters when choosing a receiver);
3) Start by contacting the most loaded clusters, so that we insure the maximum of overlap between intra and inter-cluster broadcast; (BottomUp)

The new approach (cluster of clusters)

Take into consideration 3 factors

1) We need to increase the size of set A with clusters, in the quickest possible way;
2) Availability of numerous senders give us more chance to perform next communication in a better way, since we have more choices to consider (advantage to communication-effcient clusters when choosing a receiver);
3) Start by contacting the most loaded clusters, so that we insure the maximum of overlap between intra and inter-cluster broadcast; (BottomUp)

How to choose the factor to satisfy?

The new approach (cluster of clusters)

Take into consideration 3 factors

Choose the cluster that satisfies one factor and behaves well with the other ones, or at least does not violate them strongly.

1) compute set
   $E_1 = \min_{P_i \in A}(RT_i + g_{ij}(m) + L_{ij})/P_j \in B$
2) compute set
   $E_2 = \min_{P_i \in A}(RT_i + g_{ij}(m) + L_{ij} + F_j)/P_j \in B$
3) compute set
   $E_3 = \min_{P_i \in A}(RT_i + g_{ij}(m) + L_{ij} + T_j)/P_j \in B$

The new approach (cluster of clusters)

Take into consideration 3 factors

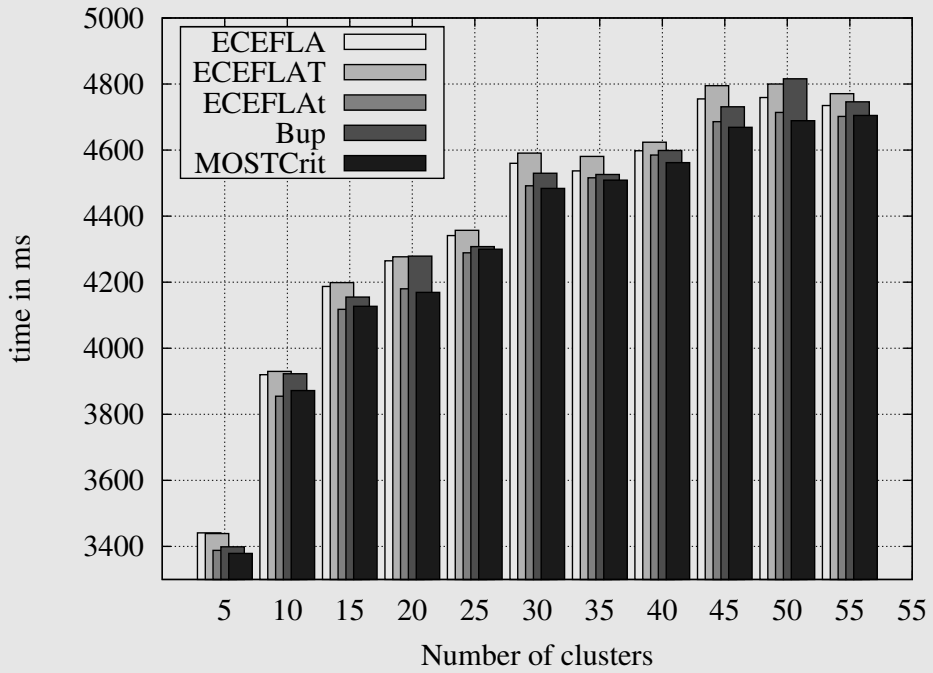
- Dispersed value in a given set $\Rightarrow$ clusters are very different according to the associated factor. Otherwise it means that clusters behave in a quite similar way. Subsequently, choosing one cluster or another one will not be decisive.
- Choose to satisfy the factor which has the associated set with the most dispersed values i.e. we compute the mean deviation of each set values and we choose to satisfy the factor having the greatest mean deviation.

Simulations

Parameters measured on Grid'5000

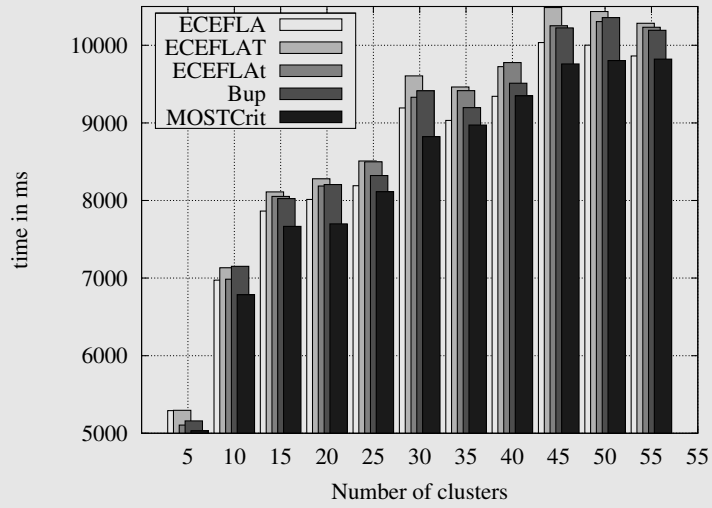
	First simulation		Second simulation		Third simulation	
param	min	max	min	max	min	max
L	1	15	5	75	10	150
g	100	600	500	3000	1000	6000
T	200	3000	40	600	20	300

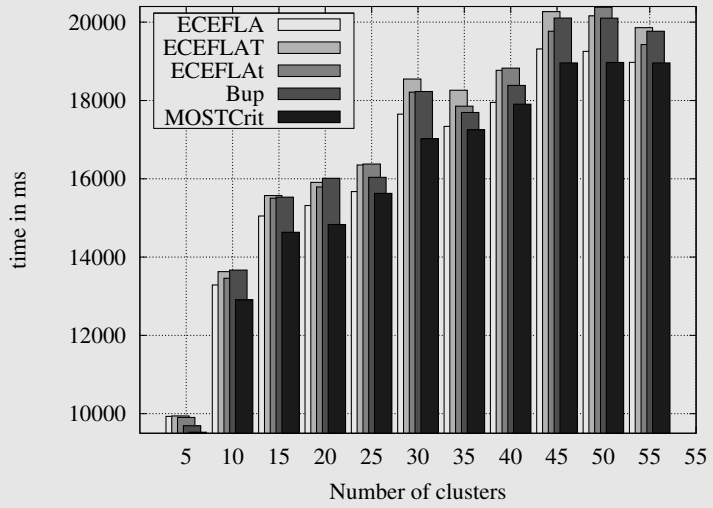
Table: Grid'5000 settings



Simulations

2nd and 3rd simulations





Experiments

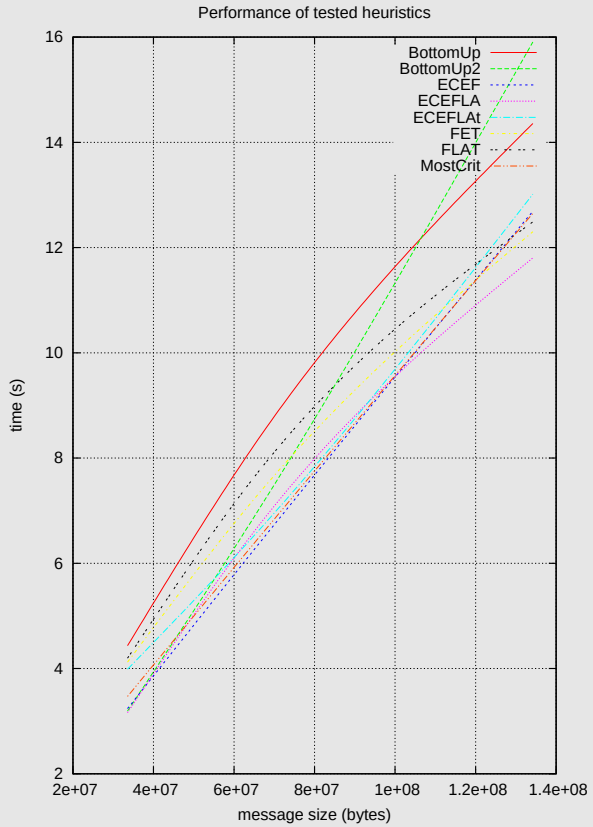
Grid'5000: a large scale instrument

Web site: https://www.grid5000.org

Experiments

Grid'5000: a large scale instrument

Experiment on 3 sites: Nancy, Rennes, Sophia-Antipolis



Conclusion

From heuristics to large scale experiments

- Contrarily to previous works, we do not attempt to combine 'factors' → apply only one factor at each iteration;
- Simulations and experiments showed that the approach is as good as the best 'heuristic' taken in a separate way (no degradation in performance) ; Different contexts for the executions;
- Future work: congestion modeling since the 'backbone' is shared among experiments ⇒ Online alg. (subproblem for simulations: a good model for congestion in cluster of cluster?)

A new heuristic for broadcasting in cluster of clusters – AOC Team –

Christophe Cérin[1], Hazem Fkaier, Luiz-Angelo Steffenel, Mohamed Jemni

[1]Université de Paris XIII, CNRS UMR 7030, France

GPC conference, Hualien, Taiwan