Efficient Load Balancing by Adaptive Bypasses for the Migration on the Internet

Yukio Hayashi

yhayashi@jaist.ac.jp

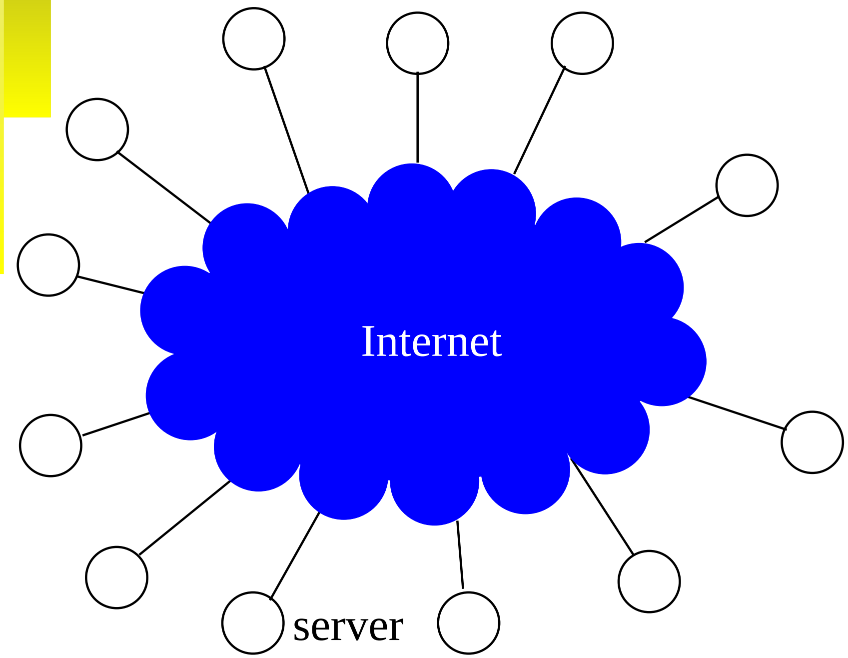
Japan Advanced Institute of Science and Technology

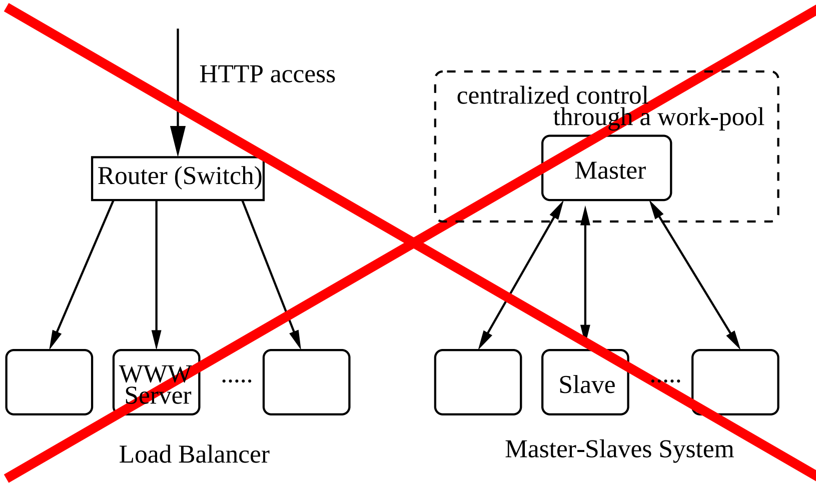
Overview

We present

- a problem setting for dynamic load balancing on the Internet as Grid,
- an adaptive method according to initial load: the optimal flow is directly obtained, the conditions of migration for the bottleneck edges are relaxed by bypasses,
- simulation results show the number of rounds of the migration is decreased by adaptive bypasses on a cactus.

1-1. Distributed System on the Internet





logically connected servers (computers) to communication partners in a general topology with routings

1-2. Load Balancing Problem

One of the important issues in distributed computing

Differences of distributed computing to parallel computing:
loose coupling, independence of process elements (divisible load), and heterogeneity (network, servers) [V.S. Sunderam & G.A. Geist ’99]

To minimize the message time-complexity in load balancing, we consider to

- avoid wasteful communications and migrations,
- based on locally asynchronous processes,
- which are dominant than the communication over heads.

1-3. Popular Approaches

Two phase methods: calculation of the optimal flow + migration of load [C. Xu & F.C.M. Lau '97]

- Dimension exchange method
  suitable for parallel machines on a hypercube structure with one-port communication

- Diffusion method $\Leftarrow$ focussed
  suitable for asynchronous processing with multi-port communication in a general topology

Others: initiation methods or round-robin, simple but ad hoc (no guarantee the global balancing)

$\Rightarrow$ We aim to globally balance the load fast by the least amount of migration and communication.

2-1. New Problem Setting on the Internet

Definitions

Load: the ratio of number of processes in ready state to the performance, denoted by $f(u)$, $u \in V$. We assume that the performances of servers are the same.

Cost: the instability of traffic (e.g. fluctuation measured by ping), denoted by $1/w_e$, $e \in E$.

server $\leftrightarrow$ vertex, communication $\leftrightarrow$ edge, on (V, E)

| Characteristic of the Internet | We don't consider the weights of communication efficiency for data transfer speed with delay: unstable and indefinite.

2-2. Diffusion Equation

Let us consider a discrete Laplacian L, the operated u-th element is

$$Lf(u) = -\sum_{v \sim u} w_e(f(v) - f(u)),$$

The DF methods essentially result in solving $\frac{\partial \mathbf{f}}{\partial t} = -L\mathbf{f}$, the flow is determined by the differences with w_e.
At the k-th iteration, the difference equation (FOS) is

$$\mathbf{f}^k = (I - \Delta t L)\mathbf{f}^{k-1} = \underbrace{F \times \ldots \times F}_{k} \mathbf{f}^0,$$

$F \stackrel{\text{def}}{=} I - \Delta t L$, the step-width satisfies $1 \leq \Delta t \times \sum_{v \sim u} w_e$.

However, the asymptotical convergence is very slow.

2-3. Finite Convergent OPS

Optimal Polynomial Scheme [R. Diekmann et al. '99]

$p_0(t) = 1$, $p_1(t) = \frac{1}{\gamma_1}[(\alpha_1 - t)p_0(t)]$,

$p_k(t) = \frac{1}{\gamma_k}[(\alpha_k - t)p_{k-1}(t) - \beta_k p_{k-2}(t)], \quad (k \geq 2)$,

$$\mathbf{f}^k = p_k(F)\mathbf{f}^0 = \frac{1}{\gamma_k}\left[\alpha_k \mathbf{f}^{k-1} - F\mathbf{f}^{k-1} - \beta_k \mathbf{f}^{k-2}\right].$$

OPS is including FOS, SOS, Chebyshev schemes, and established for parallel machines, but,

- the calculation of all eigenvalues of L is necessary in advance, to set the parameters α_k, β_k, and γ_k,
- there exists a problem for the ordering of migrations in cycles, after the calculation of flow.

$\Rightarrow$ not suitable for distributed systems

2-4. Equivalent QP Problem

The difference equation: $\mathbf{f}^k = (I - \Delta t L)\mathbf{f}^{k-1}$

$\Downarrow$ is equivalent to the following

QP problem [Y.F. Hu & R.J. Blake '99]:

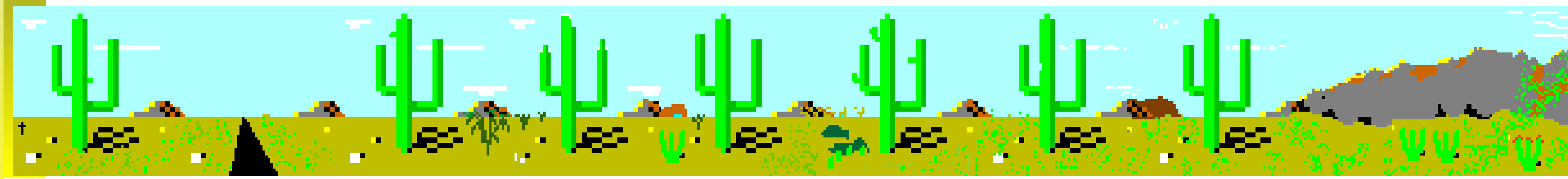
$$\begin{array}{ll} \min & \frac{1}{2}\mathbf{z}^T W^{-1}\mathbf{z}, \\ s.t. & B\mathbf{z} = \mathbf{f}^0 - \bar{\mathbf{f}}, \end{array}$$

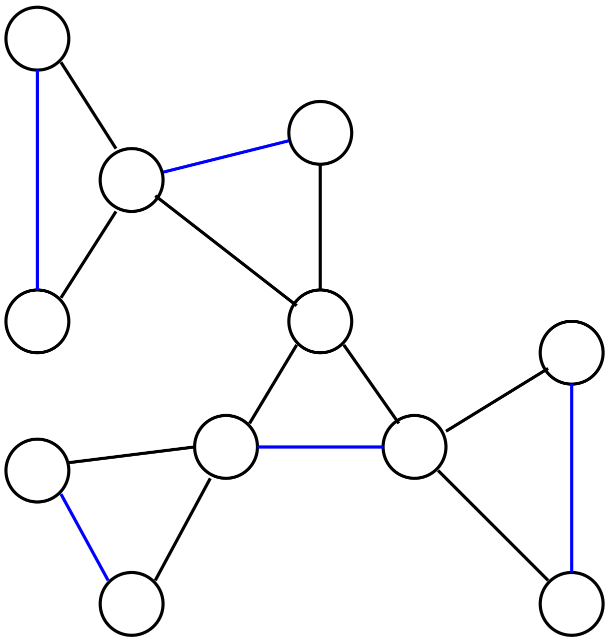
where, $W \stackrel{\text{def}}{=} diag(w_e)$, $\bar{f} \stackrel{\text{def}}{=} \frac{\sum_{u \in V} f(u)}{|V|}$ balancing solution, B incidence matrix, $\mathbf{z}$ flow vector for the migration.

$\Rightarrow$ adaptive method: calculation of the optimal flow by efficient message passing on a tree (without cycles), and bypasses of migration for the bottleneck edges on a cactus according to initial load.

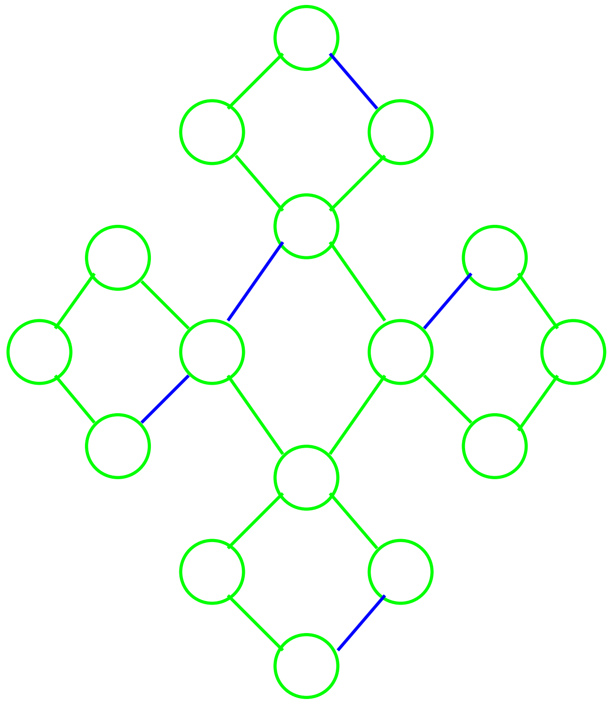
3. Adaptively Constructed Cactus

3-1. Cactus Structure: similar to plants in desert



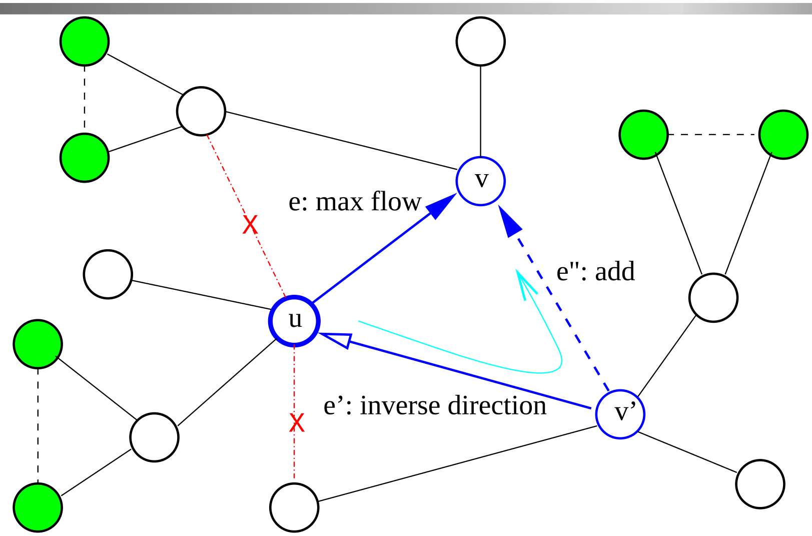


Ternary Cactus



Quaternary Cactus

3-2. Algorithm on a Cactus



At a trigger with heavy load, processes are initiated.

1. Calculate the flow z_e on the MST by applying TWA.
2. Find a bottleneck edge e, the pair e', and the candidate e'', mutually exclude the candidate of bypass e''. Calculate the modified flow $z_e - \Delta z_{opt}$, $z_{e'} - \Delta z_{opt}$, Δz_{opt} for the fixed bypass.
3. Asynchronously migrate it in locally distributed manner.

3-3. Efficient Message Passing

Tree Walking Algorithm for calculating the optimal flow

1. Accumulate the total load $\sum_{u \in V} f(u)$ from leaves to the root.
2. Broadcast the balancing solution $\bar{f}$ from the root to leaves.
3. Calculate the flow from leaves to the root.

   For a leaf u $\quad z_e = f(u) - \bar{f}$,

   where $e \in E$ is an edge to the parent of $u \in V$.

   For others $\quad z_{e'} = f(v) - \bar{f} + \sum_e z_e$,

   where $e' \in E$ is an edge to the parent of $v \in V$, and $\{e\}$ in the summation is a set of edges from the children of v.

These processes are message-driven.

3-4. Bypasses of the Bottleneck

To solve the QP problem equivalent to DF method,

variation for an extended cactus: $\min \ \frac{1}{2}\mathbf{z}^T W^{-1}\mathbf{z}$

invariant balancing due to bypass: $s.t. \ B\mathbf{z} = \mathbf{f}^0 - \bar{\mathbf{f}}$

$\Downarrow$ since each cycle is independent, no relations

For a bottleneck $e = \arg\ \max_{e \in E_u}\{|z_e|^2/w_e\}$,

$$\delta C(\Delta z) \stackrel{\text{def}}{=} \frac{(z_e - \Delta z)^2}{w_e} + \frac{(z_{e'} - \Delta z)^2}{w_{e'}} + \frac{\Delta z^2}{w_{e''}} - \left(\frac{z_e^2}{w_e} + \frac{z_{e'}^2}{w_{e'}} \right),$$

e: $u \to v$, the pair e': $u \leftarrow v'$, and the bypass e'': $v' \to v$.

At the extreme point $\frac{\partial(\delta C)}{\partial(\Delta z)} = 0$, we can derive

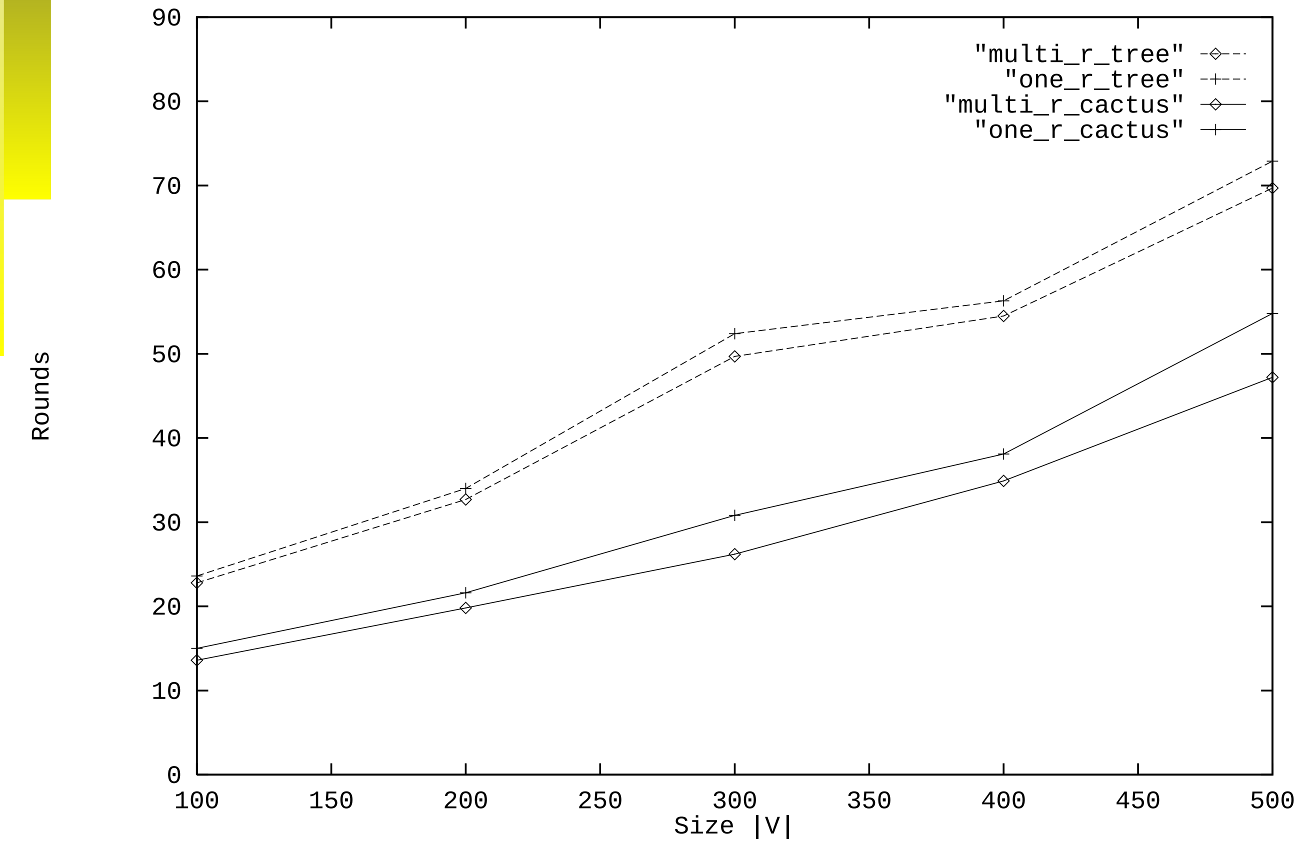
$\Delta z_{opt} = \frac{w_{e'} w_{e''} z_e + w_e w_{e''} z_{e'}}{w_{e'} w_{e''} + w_e w_{e''} + w_e w_{e'}} > 0$, and $\delta C(\Delta z_{opt}) < 0$: the cost is always decreased by adding bypasses.

3-5. Practically Better Ternary Cactus

The ternary is practically better in the reasons.

- The mutual exclusion is restricted in the alternative combination of triangles. If we consider longer cycles, it may be intractable that many edges are complicatedly related.
- Each server can directly communicate to the nearest-neighbors. While, for longer cycles, it must pass the information z_e or w_e through intermediators.
- Both ends of a bypass edge are probably close in the geographical locations.

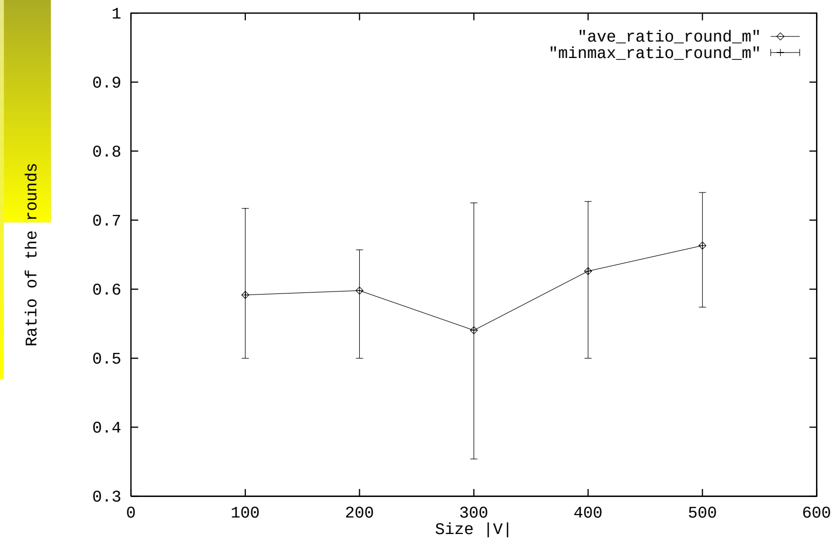
4-1. Average Rounds of Migration vs Network Size $|V|$

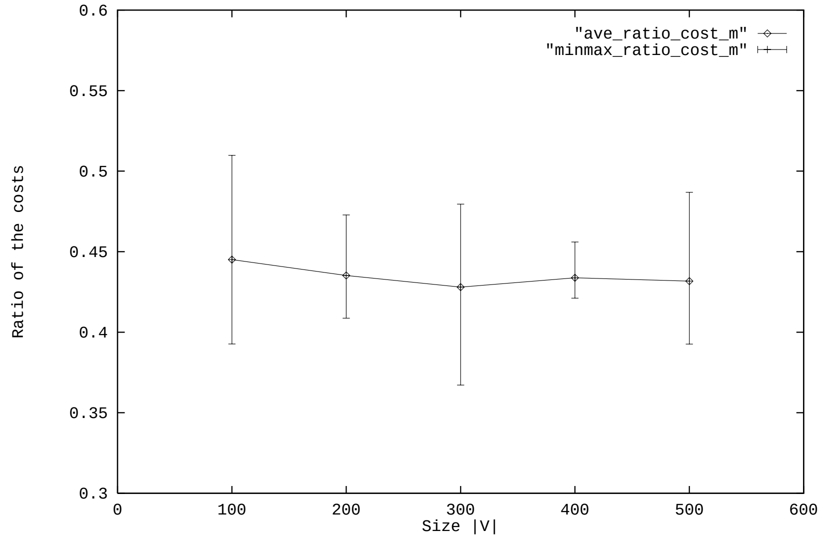


the solid and dashed lines: a cactus and the MST, multi-ports $\diamond$ is 10 % improved than the one-port $+$.

$\Rightarrow$ The results for cacti are better.

4-2. Ratio of the Rounds and Cost on a Cactus/MST (multi-ports)





the max, min, and average rounds (left) and the costs (right) in 10 trials.

$\Rightarrow$ The ratio of rounds is decreased under 2/3, and the cost w.r.t the traffic instability is about the half.

5. Conclusion

Load balancing for servers on the Internet as Grid.

1. For a QP problem equivalent to the DF method, we have proposed an adaptive method: the optimal flow is obtained by using variational computations, and the conditions of migration for the bottleneck edges are relaxed by the bypasses on a cactus.
2. We have presented it as a distributed algorithm based on only local communication and asynchronous processes.
3. As underestimations, simulation results have shown the num. of rounds for migration are decreased under 2/3 for the MST, and the cost w.r.t the traffic instability is about the half.