Some Properties of Pareto Sets in Load balancing

Corinne Touati, Atsushi Inoie, Hisao Kameda
University of Tsukuba - Graduate School of Systems and Information Engineering - Japan
$\{corinne, inoie, kameda\}@osdp.cs.tsukuba.ac.jp$

Abstract

A very important concept in optimization of computer and communication systems is the Pareto set which is the set of the optimal solutions or points of the system in question. Every Pareto optimal solution can have its own interest and significance. In this paper, we obtain several analytical results on the Pareto set of a simple distributed system with two flows arriving to two servers where each of the flow can be rerouted to the other server, inducing an additional fixed delay. We consider, as performance measure, the total delay experienced by each flow. The model leads to a non-convex optimization and its analysis does not happen to be so simple and shows various counter-intuitive behaviors, in spite of the seeming simplicity of the model. We further give some insights into the relation between the Pareto set and the set of achievable delays, and properties of the latter.

Keywords — Load balancing, Pareto optimality, non-convex optimization, distributed computer systems.

1 Introduction

Optimal load balancing among computers in a distributed system is considered to be a means of effectively utilizing resources among users. We may consider innumerably many objectives of optimization of load balancing, for example, minimizing the weighted sum of the cost for each user with different values of the weights. Each objective may correspond to a Pareto optimum. There exists no other state where all users have better benefits simultaneously than in a Pareto optimum situation. Thus, we may not have any absolute preference among Pareto optima.

In this paper, we derive some properties on the Pareto set with respect to a simple model of a distributed system that consists of two servers shared by two users or two sets of users (Kameda et al. (2000), Tantawi and Towsley (1985), Kameda et al. (1997)). We consider as the utilities of the users, the delays (or the mean response times) that the users observe. We first note that the delay functions are not convex and, then, we anticipate that optimization problems in load balancing may be much more complicated than the standard convex problems (see, for example, Aubin (1998)).

As the forwarding process induce some additional delay, one may expect the efficient solutions to be characterized by a non-mutual forwarding. In the studied systems, however, we show that counter-intuitive scenarios may occur. Even more surprisingly, an equilibrium in which the whole traffic of one user (let us say player one) is forwarded and in which player two also forwards a part of its traffic can be Pareto optimal.

We obtain a number of further analytical results. We show that the Pareto set achieved by the optimization of the weighted sum of the delay for each user consists of up to four sections of curves, of which we give an analytical formulation. Furthermore, in a special case where the forwarding cost is negligible, the remaining section of the Pareto set is a segment of line. We also obtain results on the continuity and bounds of the Pareto set.

After briefly presenting the studied system (Section 2), we present a study of the achievable set, that is to say the set of pairs of delays (Section 3). We then focus on its lower left border, that is to say on the Pareto set itself. The last section presents some additional properties observed via numerical studies.

2 Model and assumptions

We consider a distributed computer system consisting of two servers (computers), 1 and 2, with two flows of demands ϕ_1 and ϕ_2 arriving from users 1 and 2 at servers 1 and 2, respectively (Kameda *et al.* 2000). Let a fraction x_i ($0 \leq x_i \leq \phi_i$) of a flow of jobs be forwarded from server i to the other server j ($\neq i$) as depicted in Figure 1.

Denote by $\overrightarrow{x}$ the vector (x_1, x_2) and by β_1 and β_2, respectively, the resulting loads on nodes 1 and 2. Then

$$\beta_i = \phi_i - x_i + x_j, \quad i, j = 1, 2 \ (i \neq j).$$

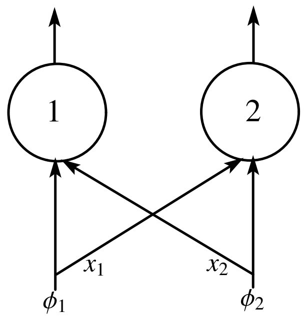

Figure 1: Load balancing in a distributed system consisting of two servers

We assume that the processing time at server i for the load of rate β_i is given by $\dfrac{1}{\mu_i - \beta_i}$. For simplicity, we suppose that forwarding a job induces a fixed delay t. Therefore, the cost of user i, that is, the delay of each flow arriving from the user i, can be written

$$T_i(\overrightarrow{x}) = \frac{1}{\phi_i}\left[\frac{\phi_i - x_i}{\mu_i - \phi_i + x_i - x_j} + x_i(t + \frac{1}{\mu_j - \phi_j + x_j - x_i})\right]. \tag{1}$$

Set of constraints We denote by $C_{\overrightarrow{x}}$ the set of constraints on (x_1, x_2). The forwarding rates should be positive or null and bounded by the arrival flow: $0 \leq x_i \leq \phi_i$. Also, the service time at each server should not grow without bounds: $\mu_i - \phi_i + x_i - x_j > 0, \{i, j\} \in [1, 2], i \neq j$. Therefore the necessary condition on the server mean service time and the arrival rate for the system to be feasible is

$$\mu_1 + \mu_2 > \phi_1 + \phi_2. \tag{2}$$

and the difference of the forwarding rates must satisfy $\phi_1 - \mu_1 < x_1 - x_2 < \mu_2 - \phi_2$. Finally,

$$C_{\overrightarrow{x}} = \{\overrightarrow{x} \mid 0 \leq x_i \leq \phi_i, i = 1, 2 \text{ and } \phi_1 - \mu_1 < x_1 - x_2 < \mu_2 - \phi_2\}.$$

Clearly, $C_{\overrightarrow{x}}$ is a convex set. The set of points $(T_1(\overrightarrow{x}), T_2(\overrightarrow{x}))$ with $\overrightarrow{x} \in C_{\overrightarrow{x}}$ is the *achievable set*. Clearly, in the plane T_1, T_2 this set is included in the first quadrant (i.e. $\forall \overrightarrow{x} \in C_{\overrightarrow{x}}, T_1(\overrightarrow{x}) > 0$ and $T_2(\overrightarrow{x}) > 0$).

Pareto set We recall that the Pareto set $\mathbf{\Pi}$ is the set of achievable pairs (T_1, T_2) such that if $(T_1^\bullet, T_2^\bullet)$ is another possible pair then $T_1^\bullet < T_1$ implies $T_2^\bullet > T_2$.

$$\mathbf{\Pi} = \{(T_1(\overrightarrow{x}), T_2(\overrightarrow{x})) \mid \overrightarrow{x} \in C_{\overrightarrow{x}}, \forall \overrightarrow{y} \in C_{\overrightarrow{x}},\ T_i(\overrightarrow{y}) < T_i(\overrightarrow{x})\ (i \in \{1, 2\}) \Rightarrow T_j(\overrightarrow{y}) > T_j(\overrightarrow{x})\ (j \neq i).\}$$

The Pareto set is therefore included in the lower left border of the achievable set, which is why we also refer to it as *Pareto border*. Studying the Pareto set is of interest since it represents the points that cannot be dominated in terms of performances. Therefore load balancing algorithms naturally (try to) reach Pareto efficient equilibria.

Since the forwarding process induce some additional delay (proportional to the forwarding time t), one can expect the Pareto set to be formed by the vectors $\overrightarrow{x}$ such that $x_1 x_2 = 0$. This result can actually be denied by simple examples. Consider for instance a system with: $\phi_1 = 0.7,\ \phi_2 = 0.9,\ \mu_1 = 1,\ \mu_2 = 1.2,\ t = 0.001$. Then $(T_1(0, 0), T_2(0, 0)) \approx (3.3333, 3.3333)$ while $(T_1(0.39, 0.4), T_2(0.39, 0.4)) \approx (3.3249, 3.3251)$. Therefore $(0, 0)$ is not a Pareto optimal point. We will see in the paper that not only some points satisfying $x_1 x_2 = 0$ are Pareto inefficient, but also, and more surprisingly, that some optimal points do not satisfy $x_1 x_2 = 0$.

We first study, in the next section, the set of achievable pairs of delays.

3 Achievable set

Our study relies on the following equivalent system:

Proposition 1 (Equivalent system) *Note that the denominator of the delay functions depends only on the difference of the forwarding rates $x_1 - x_2$. We then make the natural change of variables $u = x_1 - x_2$, $v = x_1 + x_2$. The delay* *functions become:*

$$\begin{cases} \phi_1 T_1(\overrightarrow{x}) = \dfrac{\phi_1 - (u+v)/2}{\mu_1 - \phi_1 + u} + \dfrac{u+v}{2}\,(t + \dfrac{1}{\mu_2 - \phi_2 - u}) \\ \phi_2 T_2(\overrightarrow{x}) = \dfrac{\phi_2 - (v-u)/2}{\mu_2 - \phi_2 - u} + \dfrac{v-u}{2}\,(t + \dfrac{1}{\mu_1 - \phi_1 + u}) \end{cases} \tag{3}$$

and the set of constraints is:

$$\overrightarrow{C_{uv}} = \{(u,v) \in \mathbb{R}^2 |\ 0 \le u+v \le 2\phi_1,\ 0 \le v-u \le 2\phi_2,\ \phi_1 - \mu_1 < u < \mu_2 - \phi_2\}. \tag{4}$$

Achievable set As the functions T_1 and T_2 are linear in v we can express T_2 as a function of T_1 and u. If $t + \dfrac{1}{\mu_2 - \phi_2 - u} - \dfrac{1}{\mu_1 - \phi_1 + u} \neq 0$ then

$$\phi_2 T_2 = -2 + \frac{\mu_2}{\mu_2 - \phi_2 - u} + \frac{\mu_1}{\mu_1 - \phi_1 + u} - \phi_1 T_1 - ut + \left[T_1 - \frac{1}{\mu_1 - \phi_1 + u}\right] \frac{2t\phi_1}{t + \dfrac{1}{\mu_2 - \phi_2 - u} - \dfrac{1}{\mu_1 - \phi_1 + u}}. \tag{5}$$

We recognize here the equation of the line in the plane T_1, T_2. We conclude that the set of achievable pairs (T_1, T_2) consists of segments of lines parameterized by T_1 and u. The limits of these segments are given by the extremum values of u (given by Eq. 4).

Extremum cases Note that the extremum cases are, from Eq. 4, $v = \max(u, -u)$ and $v = \min(2\phi_1 - u, 2\phi_2 + u)$. Furthermore:

$$\begin{array}{ll} v = u \leftrightarrow x_2 = 0, & v = 2\phi_1 - u \leftrightarrow x_1 = \phi_1, \\ v = -u \leftrightarrow x_1 = 0, & v = 2\phi_2 + u \leftrightarrow x_2 = \phi_2. \end{array} \tag{6}$$

We provide the equations corresponding to the extremum cases, that is to say to $x_1 x_2 (\phi_1 - x_1)(\phi_2 - x_2) = 0$. We summarize in the following table the necessary condition on the parameters for each of the cases to be achievable, the set of values of the rerouting loads and the parametric equations of the corresponding graph.

Extremum case	Necessary condition (CN) and set of values (SV)	Parametric equation
$x_1 = 0$	(CN) $\phi_1 < \mu_1$ (SV) $\phi_2 - \mu_2 < x_2 < \mu_1 - \phi_1$	$\begin{cases} T_1[0, x_2] = \dfrac{1}{\mu_1 - \phi_1 - x_2} \\ T_2[0, x_2] = \dfrac{1}{\phi_2}\left(\left(t + \dfrac{1}{\mu_1 - \phi_1 - x_2}\right) x_2 + \dfrac{\phi_2 - x_2}{\mu_2 - \phi_2 + x_2}\right) \end{cases}$
$x_1 = \phi_1$	(CN) $\phi_1 < \mu_2$ (SV) $\phi_1 + \phi_2 - \mu_2 < x_2 < \mu_1$	$\begin{cases} T_1[\phi_1, x_2] = t + \dfrac{1}{\mu_2 - \phi_1 - \phi_2 + x_2} \\ T_2[\phi_1, x_2] = \dfrac{1}{\phi_2}\left(\left(t + \dfrac{1}{\mu_1 - x_2}\right) x_2 + \dfrac{\phi_2 - x_2}{\mu_2 - \phi_1 - \phi_2 + x_2}\right) \end{cases}$
$x_2 = 0$	(CN) $\phi_2 < \mu_2$ (SV) $\phi_1 - \mu_1 < x_1 < \mu_2 - \phi_2$	$\begin{cases} T_1[x_1, 0] = \dfrac{1}{\phi_1}\left((t + \dfrac{1}{\mu_2 - \phi_2 - x_1}) x_1 + \dfrac{\phi_1 - x_1}{\mu_1 - \phi_1 + x_1}\right). \\ T_2[x_1, 0] = \dfrac{1}{\mu_2 - \phi_2 - x_1} \end{cases}$
$x_2 = \phi_2$	(CN) $\phi_2 < \mu_1$ (SV) $\phi_1 + \phi_2 - \mu_1 < x_1 < \mu_2$	$\begin{cases} T_1[x_1, \phi_2] = \dfrac{1}{\phi_1}\left(\left(t + \dfrac{1}{\mu_2 - x_1}\right) x_1 + \dfrac{\phi_1 - x_1}{\mu_1 - \phi_1 - \phi_2 + x_1}\right) \\ T_2[x_1, \phi_2] = t + \dfrac{1}{\mu_1 - \phi_1 - \phi_2 + x_1} \end{cases}$

These can also be written, in the plane T_1, T_2:

$$\begin{cases} \phi_2 T_2(0, x_2) = (t + T_1(0, x_2))(\mu_1 - \phi_1 - \dfrac{1}{T_1(0, x_2)}) - 1 + \dfrac{\mu_2 T_1(0, x_2)}{T_1(0, x_2)(\mu_1 + \mu_2 - \phi_1 - \phi_2) - 1} \\ \phi_2 T_2(\phi_1, x_2) = -2 + \dfrac{\mu_1 (T_1(\phi_1, x_2) - t)}{(T_1(\phi_1, x_2) - t)(\mu_1 + \mu_2 - \phi_1 - \phi_2) - 1} + (T_1(\phi_1, x_2) - 2t)(\mu_2 - \phi_1) + \left(\dfrac{2}{T_1(\phi_1, x_2) - t} + \phi_2\right) t \\ \phi_1 T_1(x_1, 0) = (t + T_2(x_1, 0))(\mu_2 - \phi_2 - \dfrac{1}{T_2(x_1, 0)}) - 1 + \dfrac{\mu_1 T_2(x_1, 0)}{T_2(x_1, 0)(\mu_1 + \mu_2 - \phi_1 - \phi_2) - 1} \\ \phi_1 T_1(x, \phi_2) = -2 + \dfrac{\mu_2 (T_2(x, \phi_2) - t)}{(T_2(x, \phi_2) - t)(\mu_1 + \mu_2 - \phi_1 - \phi_2) - 1} + (T_2(x, \phi_2) - 2t)(\mu_1 - \phi_2) + t\left(\phi_1 + \dfrac{2}{T_2(x, \phi_2) - t}\right). \end{cases}$$

We illustrate our result on Figure 3.

Remark 1 *The reader may note that we did not study here the case where $u \to \mu_2 - \phi_2$ or $u \to \phi_1 - \mu_1$. Actually, in the two cases, if v is not extremum (as defined in Eq. 6), we note that both $T_1(\overrightarrow{x})$ and $T_2(\overrightarrow{x})$ get infinite values.*

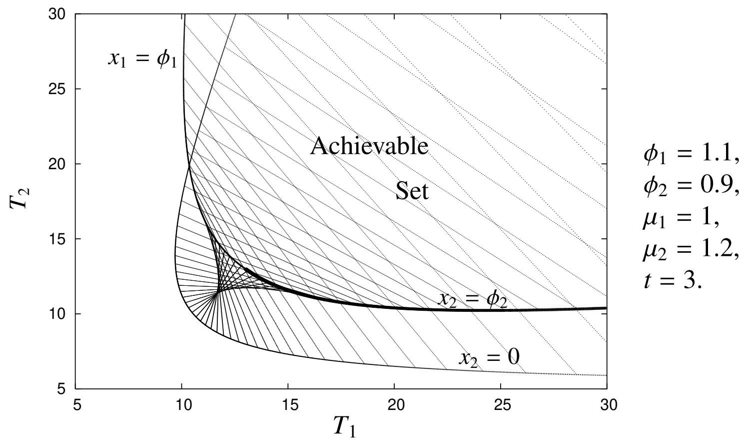

Figure 2: The achievable set is covered by segments of lines joining the extremum cases (i.e. $x_1 x_2(\phi_1 - x_1)(\phi_2 - x_2) = 0$)

Intersection points We have seen that the achievable points are formed of segments of lines. A natural question is to know how these lines intersect.

Proposition 2 *For a given pair $(T_1^\bullet, T_2^\bullet)$, the equation $(T_1(x_1, x_2), T_2(x_1, x_2)) = (T_1^\bullet, T_2^\bullet)$ admits at most five solutions.*

[Proof] From Eq. 5, if $(T_1^\bullet, T_2^\bullet)$ is given, then the problem is equivalent of finding u such that

$$\frac{\mu_2}{\mu_2 - \phi_2 - u} + \frac{\mu_1}{\mu_1 - \phi_1 + u} - ut + \left[T_1 - \frac{1}{\mu_1 - \phi_1 + u}\right] \frac{2t\phi_1}{t + \dfrac{1}{\mu_2 - \phi_2 - u} - \dfrac{1}{\mu_1 - \phi_1 + u}} = K$$

with K given ($K = \phi_1 T_1 + \phi_2 T_2 + 2$). u can then be written as the solution of a fifth-order polynomial. As T_1 is linear in v, then for a given u, the equation $T_1(u, v) = T_1^\bullet$ admits at most one solution. ■

4 Pareto border

We now focus our study on the Pareto border. Let us consider its representation in the plane T_1, T_2. The Pareto points clearly satisfies:

- (P1) For a given value of $T_1^\bullet$ there is at most one Pareto point of the form $(T_1^\bullet, T_2^\bullet)$. (Indeed, if it exists, $T_2^\bullet$ satisfies $\min\{T_2(x_1, x_2) | (x_1, x_2) \in \boldsymbol{C_x}, T_1(x_1, x_2) = T_1^\bullet\}$.)
- (P2) The Pareto border is a decreasing function (if (T_1^1, T_2^1) and (T_1^2, T_2^2) are Pareto optimal with $T_1^1 < T_1^2$ then $T_2^2 < T_2^1$).

Motivated by (P1), for each achievable delay T_1 we note $T_2^{\min}(T_1) = \min\{T_2 | \exists x \in \boldsymbol{C_{\vec{x}}}, T_1 = T_1(\vec{x}) \text{ and } T_2 = T_2(\vec{x})\}$.

4.1 Convexity and weighted sum optimization (WSO)

One important feature of the utility functions (here the delay functions) is convexity. Indeed, if the utilities are convex, then the Pareto border is given by the solutions of weighted sum optimization (WSO) (Aubin 1998).

Proposition 3 *The functions T_1 and T_2 are not convex.*

[Proof] We show the proposition for T_1, the equation on T_2 being similar. One can check that the determinant of the Hessian of T_1 is strictly negative :

$$\frac{\partial^2 T_1}{\partial u^2} \frac{\partial^2 T_1}{\partial v^2} - \left(\frac{\partial}{\partial v} \frac{\partial T_1}{\partial u}\right)^2 = -\frac{1}{\phi_i^2} \left[\frac{1}{2(\mu_2 - \phi_2 - u)^2} + \frac{1}{2(\mu_1 - \phi_1 + u)^2}\right]^2 < 0.$$ ■

Definition 1 *The minimization of a weighted sum of costs of users gives the following set:*

$$\text{(WSO)}\ \underline{\overrightarrow{X}} = \{argmin_{\overrightarrow{x} \in C_{\overrightarrow{x}}} \Omega_\xi(\overrightarrow{x}),\ 0 \le \xi \le 1\}\ \text{where}\ \Omega_\xi(\boldsymbol{x}) = \xi T_1(\overrightarrow{x}) + (1-\xi) T_2(\overrightarrow{x}). \tag{7}$$

Lemma 1 $\underline{\overrightarrow{X}}$ *gives Pareto optimum points.*

[Proof] Consider $\overrightarrow{x} \in \underline{\overrightarrow{X}}$ and (T_1, T_2) its associated delays. Then consider an achievable pair $(T_1^\bullet, T_2^\bullet)$ with $T_1^\bullet < T_1$. Then, as $\exists \xi,\ 0 \le \xi \le 1,\ \xi T_1 + (1-\xi)T_2$ is minimum, we can write $\xi T_1 + (1-\xi)T_2 \le \xi T_1^\bullet + (1-\xi)T_2^\bullet$. Then $0 > \xi(T^\bullet - T_1) \ge (1-\xi)(T_2 - T_2^\bullet)$ and finally $T_2^\bullet > T_2$. ■

We can observe an interesting result:

Proposition 4 (WSO) *The solutions of the WSO problem correspond to extremum cases, that is to say that they satisfy*

$$x_1 x_2 (\phi_2 - x_2)(\phi_1 - x_1) = 0.$$

[Proof] Note that T_1 and T_2 are linear functions of v. Therefore, for any value of ξ, the function $(u, v) \mapsto \xi T_1(\overrightarrow{x}) + (1-\xi)T_2(\overrightarrow{x})$ is linear in v. Its coefficient is $t + (1-2\xi)\left[(\mu_1 - \phi_1 + u)^{-1} - (\mu_2 - \phi_2 - u)^{-1}\right]$.

If this coefficient is non null, then the minimum value of $\xi T_1(u, v) + (1-\xi)T_2(u, v)$ corresponds therefore to either a minimal or a maximal value of v. Let us now examine the case where $t + (1-2\xi)\left[(\mu_1 - \phi_1 + u)^{-1} - (\mu_2 - \phi_2 - u)^{-1}\right] = 0$. Then $\xi T_1(u, v) + (1-\xi)T_2(u, v)$

$$\begin{aligned}
&= 4(-1+\xi)\xi - \frac{\xi(2(-1+\xi)\mu_1 + \phi_1 - 2\xi\phi_1)}{\mu_1 - \phi_1 + u} + \frac{(-1+\xi)(2\xi(\mu_2 - \phi_2) + \phi_2)}{-\mu_2 + \phi_2 + u} \\
&= (\xi - 1)\left(4\xi + \frac{t(2\xi(\mu_2 - \phi_2) + \phi_2)}{1 - 2\xi}\right) - \frac{\xi(-2\mu_1 + 2\mu_2 + \phi_1 - 3\phi_2) + \phi_2 + 2\xi^2(\mu_1 - \mu_2 - \phi_1 + \phi_2)}{\mu_1 - \phi_1 + u}.
\end{aligned}$$

This is a monotonous function of u and therefore reaches its minimum value when u is extremum. ■

We therefore observe that while the results obtained for $x_1 x_2 = 0$ are quite intuitive, a scenario in which all the traffic is rerouted can also be Pareto optimal.

4.2 Continuity and bounds

Any achievable delay is (strictly) positive. Therefore, for given values of ϕ_1, ϕ_2, μ_1 and μ_2, there exists $T_1^{\min} \ge 0$ and $T_2^{\min} \ge 0$ such that $\forall \overrightarrow{x} \in C_{\overrightarrow{x}}, T_1(\overrightarrow{x}) \ge T_1^{\min}$ and $T_2(\overrightarrow{x}) \ge T_2^{\min}$. Whether the Pareto set is (upper) bounded actually depends on the relative values of ϕ_1, ϕ_2, μ_1 and μ_2. The following propositions are consequences of Lemma 2:

Lemma 2 *If* $T_2^{min}(T_1) \to_{+\infty} +\infty$ *there exists* T_1^{max} *such that* $(T_1^\bullet, T_2^\bullet) \in \mathbf{\Pi} \Rightarrow T_1^\bullet < T_1^{max}$.

[Proof] Let $\overrightarrow{x^\diamond} \in C_{\overrightarrow{x}}$. As $T_2^{\min}(T_1) \to_{+\infty} +\infty$, then $\exists T_1^*, \forall T_1 \ge T_1^*, T_2^{\min}(T_1) \ge T_2(\overrightarrow{x^\diamond})$. Therefore, $\forall T_1 \ge \max(T_1^*, T_1(\overrightarrow{x^\diamond}))$, $\nexists \overrightarrow{x} \in C_{\overrightarrow{x}}, T_1(\overrightarrow{x}) = T_1$ and $(T_1(\overrightarrow{x}), T_2(\overrightarrow{x})) \in \mathbf{\Pi}$. ■

Proposition 5 *Let* $i \in \{1, 2\}$. *For given values of* $\phi_1, \phi_2, \mu_1, \mu_2$ *and t, such that* $\phi_i < \min(\mu_1, \mu_2)$, *there exists* T_i^{max} *such that* $\{\overrightarrow{x} \in C_{\overrightarrow{x}}$ *and* $(T_1(\overrightarrow{x}), T_2(\overrightarrow{x})) \in \mathbf{\Pi}\} \Rightarrow T_i(\overrightarrow{x}) \le T_i^{max}$. *We say that the Pareto set is* ***bounded in*** T_i.

[Proof] By symmetry, we prove the proposition in the case where $i = 1$. If T_1 grows to infinity, then, from Eq. 3, either $u \to \phi_1 - \mu_1$ or $u \to \mu_2 - \phi_2$. If $u \to \phi_1 - \mu_1$ then T_2 also grows to infinity unless $u \to v$ which implies $\phi_1 \ge \mu_1$ from Eq. 4. Similarly, if $u \to \mu_2 - \phi_2$ then T_2 grows to infinity unless $v \to 2\phi_2 + u$ which implies that $\mu_2 \le \phi_1$. Therefore $\phi_2 < \min(\mu_1, \mu_2) \Rightarrow T_2^{\min}(T_1) \to_{+\infty} +\infty$. Lemma 2 leads to the conclusion. ■

Proposition 6 *Let* $i, j \in \{1, 2\}, i \ne j$. *If* $\mu_i \le \phi_i$, *then, there exists* T_i^* *such that the function* T_j^{min} *is decreasing in* $[T_i^*; +\infty[$ *and its limit is* $1/(\mu_1 + \mu_2 - \phi_1 - \phi_2)$. *As* $1/(\mu_1 + \mu_2 - \phi_1 - \phi_2)$ *is a lower bound on* T_j, *we conclude that* $\nexists T_i^*, (T_1, T_2) \in \mathbf{\Pi} \Rightarrow T_i \le T_i^*$: *the Pareto set is* ***unbounded in*** T_i.

If $\mu_j \le \phi_i$ *(and* $\phi_i \le \mu_i$*) then* $\lim_{T_i \to +\infty} T_j^{min} = t + (\mu_1 + \mu_2 - \phi_1 - \phi_2)^{-1}$.

[Proof] Note from the proof of Prop. 5 that, when T_1 becomes large, T_2 becomes large as well unless x_2 equals either 0 or ϕ_2, thus the limit value of T_2. As T_1 and T_2 are linear in v, the pair (u, v) minimizing T_1 (respectively T_2) will correspond to a minimal or maximal value of v. Therefore the minimal value of T_2 is achieved in one of the extremum cases. By looking at their parametric equation, we conclude that $1/(\mu_1 + \mu_2 - \phi_1 - \phi_2)$ is the minimal value achievable by T_2. ■

Numerical results tend to show that if $\mu_j \le \phi_i$ (and $\phi_i < \mu_i$), then there exists $t^\bullet > 0$ such that $\mathbf{\Pi}$ is bounded in T_i for any $t \ge t^\bullet$ and unbounded otherwise.

We illustrate these results on some numerical experiments, in which $\phi_1 = \phi_2 = 2, \mu_1 = 1.5$ and $\mu_2 = 3$. As $\phi_1 > \mu_1$ then *the Pareto set is unbounded in* T_1. As $\phi_2 > \mu_1$, it can be unbounded or not in T_2, depending on the value of t. In any case $T_1^{\min} \to_{T_2 \to +\infty} t + (\mu_1 + \mu_2 - \phi_1 - \phi_2)^{-1}$. Results show that the Pareto set is unbounded in T_2 for low values of t and bounded otherwise.

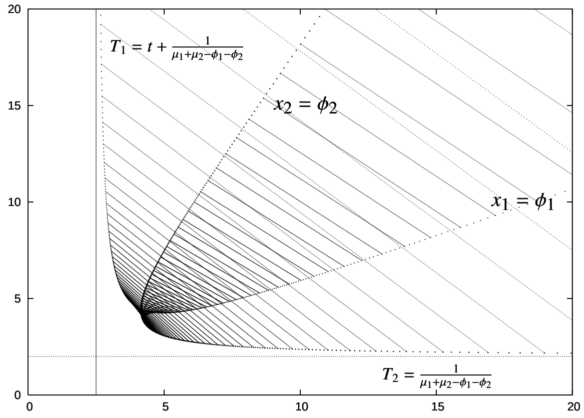

Figure 3: The Pareto set is unbounded in T_1 and T_2 ($t = 0.5$).

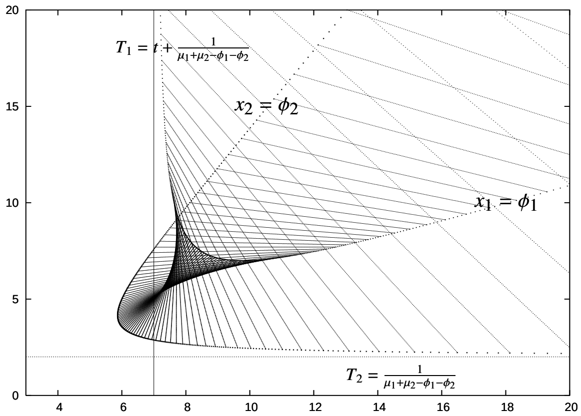

Figure 4: The Pareto set is unbounded in T_2 only ($t = 5$).

Continuity property One may at first sight believe that the Pareto set is continuous. However, this is not necessary the case as shown by Figure 6 (example with $\phi_1 = 0.4, \phi_2 = 0.5, \mu_1 = 1, \mu_2 = 1.2, t = 0.1$).

4.3 Case $t = 0$

We focus in this section on the case where the forwarding rate are null.

Proposition 7 *If $t = 0$, the set of achievable points coincides with a family of segments of lines of slope $-\phi_1/\phi_2$. The Pareto set is formed by the solutions of the WSO problem and a segment of the line* $\frac{1}{\phi_2}\left(-\phi_1 T_1 - 2 + \frac{(\sqrt{\mu_1} + \sqrt{\mu_2})^2}{\mu_1 + \mu_2 - \phi_1 - \phi_2}\right)$.

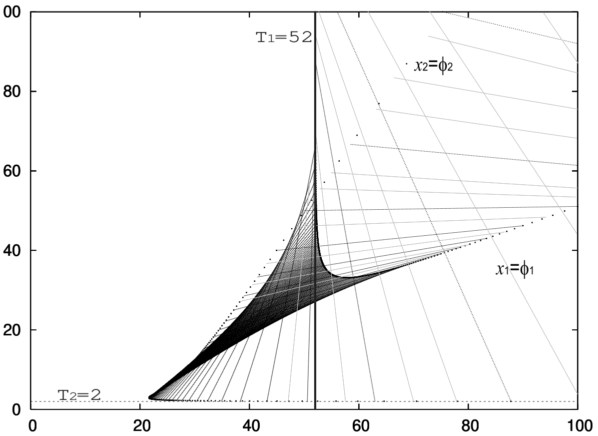

Figure 5: $t = 50$: Pareto set bounded in T_2 ($t = 50$).

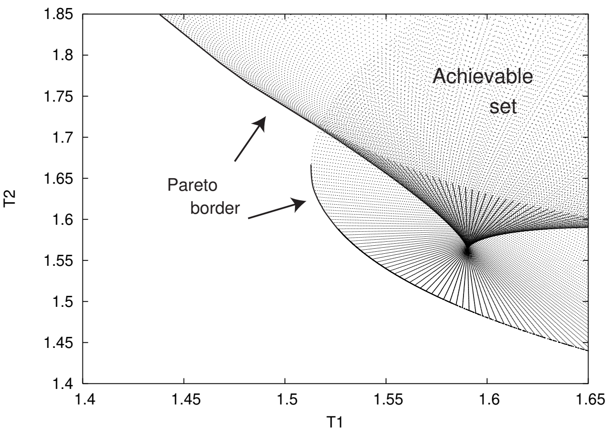

Figure 6: The Pareto set can be discontinuous

[Proof] We have seen that the achievable points consist of segments of lines. We rewrite Eq. 5 in the case $t = 0$: $T_2(\vec{x}) = \frac{1}{\phi_2}\left(-2 + \frac{\mu_1}{\mu_1 - \phi_1 + u} + \frac{\mu_2}{\mu_2 - \phi_2 - u} - \phi_1 T_1(\vec{x})\right)$. Therefore the achievable set consists of segments of lines of slope $-\phi_1/\phi_2$. Then, the Pareto set consists of the solutions given by the WSO problem and a segment of line those equation in the plane T_1, T_2 is $-\phi_1/\phi_2 T_1 + \beta$ with $\beta = \min_u \frac{1}{\phi_2}\left(-2 + \frac{\mu_1}{\mu_1 - \phi_1 + u} + \frac{\mu_2}{\mu_2 - \phi_2 - u}\right)$

We then defined f by $f : x \mapsto -\phi_1 T_1 - 2 + \mu_2/(\mu_2 - \phi_2 - x) + \mu_1/(\mu_1 - \phi_1 + x)$. It is derivable and its derivative is: $f'(x) = \mu_2(\mu_2 - \phi_2 - x)^{-2} - \mu_1(\mu_1 - \phi_1 + x)^{-2}$. The sign of f' is given by the sign of polynomial P defined by $P(x) = \mu_2(\mu_1 - \phi_1 + x)^2 - \mu_1(\mu_2 - \phi_2 - x)^2$. If $\mu_1 \neq \mu_2$ this is a second order polynomial of discriminant $\Delta = 4\mu_1\mu_2(\mu_1 + \mu_2 - \phi_1 - \phi_2)^2 > 0$. It has therefore two real roots u_1 and u_2. Since its highest coefficient is $\mu_2 - \mu_1 < 0$, then, if $\mu_2 < \mu_1$ (respectively $\mu_2 > \mu_1$), then P is decreasing on $]-\infty; u_1[\cup]u_2; +\infty[$ (resp. increasing) and increasing in $[u_1; u_2]$ (resp. decreasing). Therefore f is minimal for

$$x^* = \frac{\mu_1\phi_2 + \mu_2\phi_1 - 2\mu_1\mu_2 + \sqrt{\mu_1\mu_2}(\mu_1 + \mu_2 - \phi_1 - \phi_2)}{\mu_2 - \mu_1}.$$

If $\mu_1 = \mu_2$ then $P(x) = (\mu_1 + \mu_2 - \phi_1 - \phi_2)(2x + \phi_2 - \phi_1)$ and f is minimal for $x = (\phi_1 - \phi_2)/2$, hence the result. ■

This result is illustrated in Figures 7-8 for the case where $\phi_1 = 1, \phi_2 = 1.2, \mu_1 = 1.5, \mu_2 = 3, t = 0$.

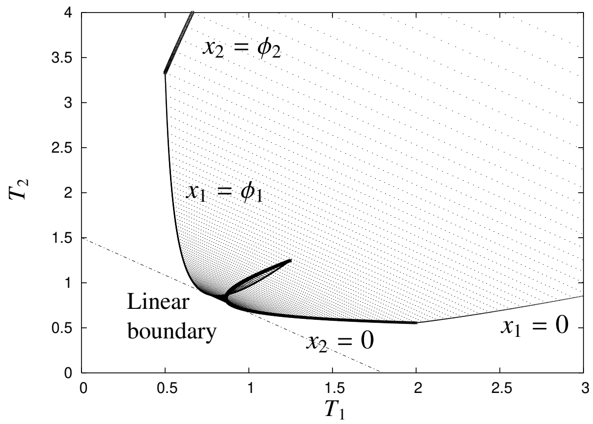

Figure 7: Case $t = 0$: the achievable set consists of parallel lines.

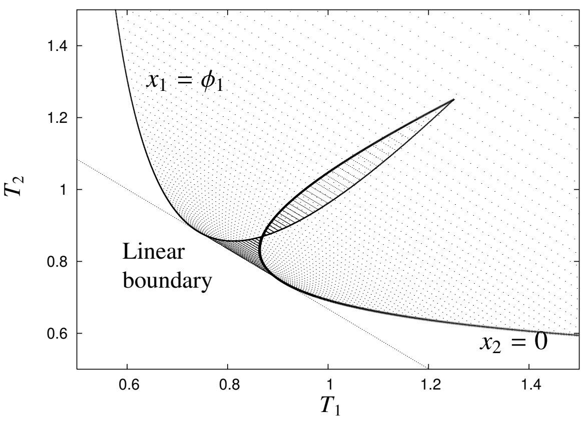

Figure 8: Case $t = 0$: the Pareto set consists of the solutions given by the WSO and a segment of line.

5 Some additional properties based on numerical experiments

Influence of the forwarding delay t [Figure 5]

Our experiments show that, as t increases, the graphs of the extremum become sharper and the Pareto set decreases. In the case where the Pareto set is bounded in T_1 and T_2 (that is to say when $\phi_i < \mu_i$, $i \in \{1, 2\}$), we observe that, for large values of t the Pareto set consists of a unique point that corresponds to the case $x_1 = x_2 = 0$.

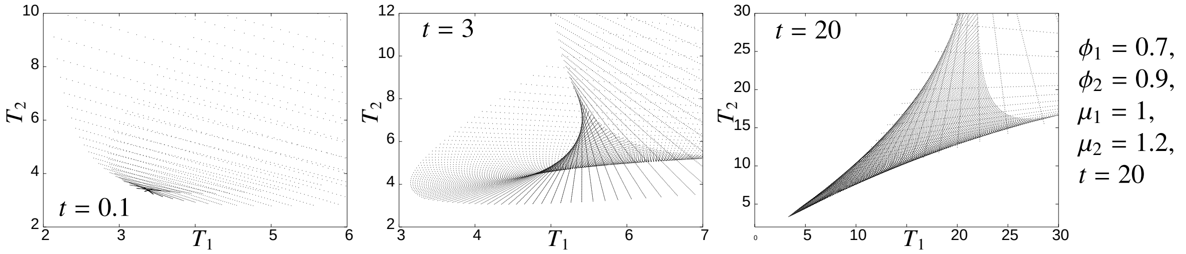

Figure 9: Evolution of the Pareto set with the forwarding delay t.

Graphs of the extremum cases [Figure 5]

If $\phi_i < \mu_i$, $i \in \{1, 2\}$, it seems that the sections of the graphs corresponding to $x_1 = 0$ and $x_2 = 0$ do not intersect each other. Instead, they join at the point $T_1(0, 0), T_2(0, 0)$ and form, together, one curve. If $\phi_i < \mu_j$, $i, j \in \{1, 2\}, i \neq j$, the same observation holds for the graphs corresponding to $x_i = \phi_i$, $i \in \{1, 2\}$.

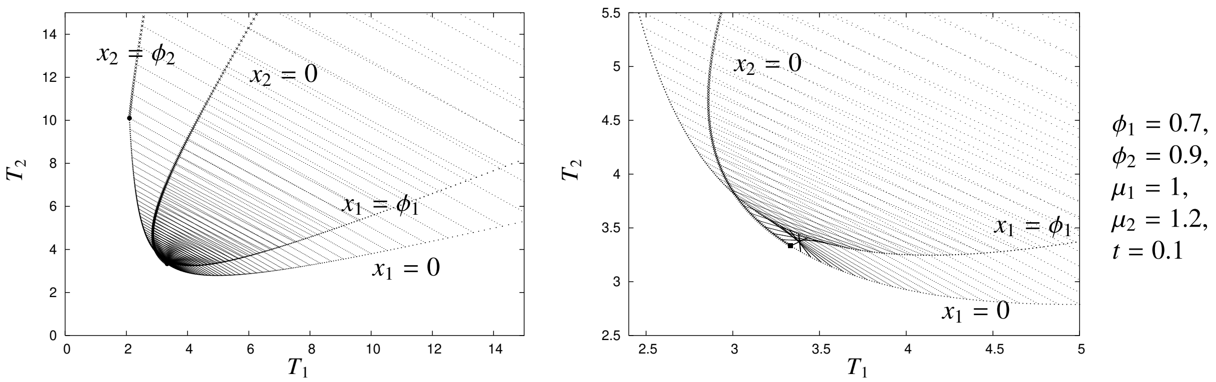

Figure 10: The graphs corresponding to $x_1 = \phi_1$ and $x_2 = \phi_2$ (respectively $x_1 = 0$ and $x_2 = 0$) do not intersect.

6 Conclusion

In this paper we have derived some properties on the Pareto set on a simple load balancing system consisting of two servers and two flows of users. We have assumed that the delay induced by forwarding a flow is fixed. We have considered as a cost function the total delay experienced by the flow.

We have given the expression and interpretation of a family of segments (or portions of lines) that coincides with the achievable set of pairs of delays. We have shown that at most five of them intersect in a given point, meaning that each achievable pair of delays can be attained by at most five strategies (i.e. choices of forwarding rate) in a given load balancing system. Furthermore, the limits of these segments are given by the (four) graphs corresponding to the minimum and maximum possible forwarding rates.

We have studied that the Pareto set, that is to say, the points that cannot be strictly dominated by other points in terms of performance. We have seen whether the Pareto border can be continuous or not, bounded or not. In particular, we have shown that, in the case where the users arrival rates are lower than the processing ones, the Pareto border is bounded. Finally, in the case of a null forwarding time, we have seen that the Pareto border consists of parts of the graphs corresponding to the extremum cases and a line of which we gave an explicit formulation. Another interesting remark is that, as the cost functions considered here are not convex, the Pareto border cannot be found from the solution of a weighted sum cost minimization. Another consequence is that the Pareto set is not necessary continuous which arises questions on optimization techniques. Through numerical experiments, we have observed that the size of the Pareto border decreases with the forwarding time and that the extremum cases actually consist of two curves.

We wish that these results may not be helpful only in the further analysis of this particular load balancing problem, but also in obtaining insight into more general resource allocation problems. In particular, we would be interested in studying the relationship between the Pareto set and the Nash equilibria which are not Pareto-optimal in general.

References

Altman, E., H. Kameda and Y. Hosokawa (2002). Nash equilibria in load balancing in distributed computer systems. *International Game Theory Review* **4**(2), 91–100.

Aubin, J-P (1998). *Optima and Equilibria: An Introduction to Nonlinear Analysis, 2nd Ed.*. Springer-Verlag. Berlin.

Kameda, H., E. Altman, T. Kozawa and Y. Hosokawa (2000). Braess-like paradoxes in distributed computer systems. *IEEE Trans. Automatic Control* **45**(9), 1687–1691.

Kameda, H., J. Li, C. Kim and Y. Zhang (1997). *Optimal Load Balancing in Distributed Computer Systems*. Springer.

Tantawi, A. N. and D. Towsley (1985). Optimal static load balancing in distributed computer systems. *J. ACM* **32**(2), 445–465.