Bornes du temps de réponse des services Web composites

Serge Haddad[*,+], **Lynda Mokdad**[**,+], **Samir Youcef**[**,+]

[*] *LSV, CNRS & ENS de Cachan*
haddad@lsv.ens-cachan.fr
[**] *LAMSADE, Université Paris-Dauphine*
{lynda.mokdad, samir.youcef}@lamsade.dauphine.fr
[+] *dans le cadre du projet ANR-06-SETI-002 Checkbound*

RÉSUMÉ. *La qualité de service (QoS) des services Web est un facteur clef de leur réussite. Ceci nécessite le développement de nouvelles méthodes afin de l'analyser. Nous proposons ici des familles de modèles majorant le temps de réponse des services Web composites pour deux types de composition : le « fork and merge » statique et aléatoire. Pour le premier cas, la complexité de résolution des modèles bornants est en $O(n\sqrt{n})$ où n est le nombre de services alors que la complexité de résolution du modèle exact est en $O(n^2)$. Pour le deuxième cas, la complexité de résolution des modèles bornants reste en $O(n\sqrt{n})$ alors que la complexité de résolution du modèle exact est en $O(n^3)$. De plus, disposer d'une famille de modèles bornants permet de choisir le modèle bornant en fonction des paramètres du modèle exact. Les résultats numériques montrent l'intérêt de notre approche en terme de complexité et de qualité de la borne.*

ABSTRACT. *The quality of service (QoS) of Web services is a key factor of their success. This requires to design new methods in order to study it. Here we propose families of upper bounding models for the response time of composite Web services for two kinds of composition: the statical and random « fork and merge ». In the first case, the complexity of bounding models belongs to $O(n\sqrt{n})$ where n is the number of called services whereas the complexity of the exact model belongs to $O(n^2)$. In the second case, the complexity of bounding models still belongs to $O(n\sqrt{n})$ whereas the complexity of the exact model belongs to $O(n^3)$. Furthermore, having a family of bounding models allows to choose the bounding model depending on the parameters of the exact model. The numerical results show the interest of our approach w.r.t. complexity and accuracy of the bound.*

MOTS-CLÉS : *Evaluation de performance, services Web, chaînes de Markov, ordre stochastique*

KEYWORDS: *Performance evaluation, Web services, Markov chain, stochastic order*

1. Introduction

Un service Web désigne essentiellement une application (ou un programme) mise à disposition sur Internet par un fournisseur de service, et accessible par les clients à travers des protocoles Internet standards (Bussler *et al.*, 2002). Les services Web dits élémentaires, tels qu'ils sont décrits par WSDL (*Web Service Description Language*), sont conceptuellement limités à des fonctionnalités relativement simples qui sont déclarées comme une collection d'opérations sans flux de contrôle. Pour certains types d'applications, il est nécessaire de combiner un ensemble de services Web élémentaires en services Web plus complexes (dits services Web agrégés ou composites) afin de répondre à des exigences plus complexes (Curbera *et al.*, 2001; Yang *et al.*, 2001).

Les services Web sont plus portables que les précédentes technologies (Bray *et al.*, 2004) mais ils soulèvent des problèmes de nature différente tels que l'adaptabilité dynamique, la qualité de service rendue. Cette dernière est définie comme étant une combinaison de plusieurs critères qui peuvent être qualitatifs (sécurité) ou quantitatifs (temps de réponse) (Menascé, 2002). Leur complexité grandissante nécessite le développement de méthodes et d'outils afin de surveiller et d'analyser leur QoS (*Quality of Service*), car une dégradation de cette dernière peut engendrer de sérieuses conséquences dont un impact économique important.

Nous nous intéressons ici au calcul du temps de réponse d'un service Web composite (SWC) dont le résultat des différents services élémentaires participant à sa composition est traité par un composant de fédération. Dans une précédente étude (Haddad *et al.*, 2008) nous avons considéré les différents patrons de contrôle supportés par BPEL (*Business Process Execution Language For Web Services*). Ici les patrons de contrôle considérés ne sont pas couverts par BPEL :

– l'invocation en parallèle d'un nombre fixe de services élémentaires et la synthèse de leur résultat par l'appelant,

– l'invocation en parallèle d'un nombre aléatoire de services élémentaires et la synthèse de leur résultat par l'appelant.

Sous l'hypothèse de services élémentaires homogènes de type markovien et d'une synthèse de résultats également markovienne, la modélisation du service composite conduit à une chaîne de Markov de taille $O(n^2)$ où n est le nombre de services invoqués. La structure particulière de cette chaîne permet d'établir des équations de récurrence qui conduisent à une résolution dont la complexité est également en $O(n^2)$. Dans les systèmes ouverts tels que les environnements pair à pair, le nombre de services élémentaires pourra atteindre des valeurs comprises entre 10^3 et 10^6 et par conséquent leur analyse exacte devient difficile voire impossible. En nous basant sur la comparaison stochastique et plus particulièrement sur la technique du couplage, nous exhibons une transformation générique de la chaîne qui garantit que le temps de réponse de la nouvelle chaîne est un majorant du temps de réponse de la chaîne initiale. Nous instancions cette transformation générique de trois façons, chacune dotée d'un paramètre « quantitatif » de transformation, aboutissant ainsi à la définition de trois familles de modèles bornants. Pour un choix approprié de ce paramètre, les systèmes

d'équation de récurrence se résolvent en espace $O(n)$ et en temps $O(n\sqrt{n})$. De plus nous montrons qu'en fonction des valeurs numériques des paramètres de la chaîne initiale, on peut toujours choisir une famille dont la borne est de bonne qualité.

Si de plus, un service peut être appelé avec une probabilité constante, alors l'étude exacte consiste à effectuer une moyenne pondérée du temps de réponse en fonction du nombre de services invoqués ce qui conduit à une résolution en temps $O(n^3)$. Dans ce cas, un usage astucieux des bornes de Chernoff permet de limiter cette étude à uniquement deux cas (le pire des cas et un « pire cas » probabiliste proche de la moyenne). Cette méthode permet de conserver la même complexité de résolution pour le calcul de bornes que dans le cas fixe.

Ce papier est organisé comme suit : dans la section 2, nous présentons quelques définitions et propriétés sur la comparaison des chaînes de Markov et les bornes de Chernoff. Dans la section 3, nous décrivons de manière formelle le modèle correspondant au patron « invocation parallèle et nombre constant de services élémentaires ». Nous présentons dans la section 4, les résultats numériques obtenus dans le cas de nombre de services élémentaires constant. Dans la section 5, nous décrivons de manière formelle le modèle correspondant au patron « invocation parallèle et nombre aléatoire de services élémentaires ». Nous présentons dans la section 6, les résultats numériques pour ce dernier cas. La section 7 conclut le papier.

2. Rappels probabilistes

Une *chaîne de Markov à temps continu* (CTMC) $\mathcal{M}$ est donnée par un espace d'états S, une matrice réelle $Q : S \times S \mapsto \mathbb{R}$ appelée générateur infinitésimal telle que (1) $\forall s \neq s' \in S\ Q[s, s'] \geq 0$ et (2) $\forall s \in S\ \sum_{s' \in S} Q[s, s'] = 0$ et une distribution initiale sur S notée $\mathcal{M}(0)$. On note $\mathcal{M}(t)$, la distribution de la chaîne au temps $t \geq 0$.

Un couplage de deux chaînes de Markov est une chaîne « produit » à valeurs dans un sous-ensemble de l'espace produit telle qu'en restreignant son observation à l'un des deux compostantes, on obtienne la chaîne correspondante.

Définition 1 *Soient $\mathcal{M} = (S, Q, \mathcal{M}(0))$ et $\mathcal{M}' = (S', Q', \mathcal{M}(0))$ deux CTMC, un couplage de $\mathcal{M}$ et $\mathcal{M}'$ est une CTMC $\mathcal{M}^* = (S^*, Q^*, \mathcal{M}^*(0))$ telle que :*

- $S^* \subseteq S \times S'$;
- $\forall s \in S,\ \mathcal{M}(0)[s] = \sum_{(s,s') \in S^*} \mathcal{M}^*(0)[s, s']$ *et*
  $\forall s' \in S',\ \mathcal{M}'(0)[s'] = \sum_{(s,s') \in S^*} \mathcal{M}^*(0)[s, s']$
- $\forall s \neq s_1 \in S,\ \forall (s, s') \in S^*,\ Q[s, s_1] = \sum_{(s,s'),(s_1,s_1') \in S^*} Q^*[(s, s'), (s_1, s_1')]$ *et*
  $\forall s' \neq s_1' \in S',\ \forall (s, s') \in S^*,\ Q'[s', s_1'] = \sum_{(s,s'),(s_1,s_1') \in S^*} Q^*[(s, s'), (s_1, s_1')]$

La technique du couplage est liée aux propriétés de l'espace d'états S^*. La proposition suivante (dont la preuve découle immédiatement de la définition du couplage) sera suffisante pour établir nos bornes sur le temps de service.

Proposition 1 *Soit $\mathcal{M}^*$ un couplage de $\mathcal{M}$ et de $\mathcal{M}'$, soient $s_f \in S$ et $s'_f \in S'$ tels que $\forall (s, s') \in S^*\ s' = s'_f \Rightarrow s = s_f$. Alors, en notant $T_{\mathcal{M}}(s_f)$ (resp. $T_{\mathcal{M}'}(s'_f)$) le temps d'atteinte moyen de s_f (resp. s'_f) dans $\mathcal{M}$ (resp. $\mathcal{M}'$), on a :*

$$T_{\mathcal{M}}(s_f) \leq T_{\mathcal{M}'}(s'_f)$$

Les bornes de Chernoff fournissent une borne précise sur la probabilité de déviation de la moyenne d'une variable aléatoire donnée sous la forme d'une somme de variables aléatoires à valeurs dans $\{0, 1\}$ et indépendantes.

Proposition 2 *(Bornes de Chernoff) Soient $X_1, ..., X_n$ des v.a indépendantes à valeurs dans $\{0, 1\}$ telles que $P(X_i = 1) = p_i$, $X \equiv \sum_{i \leq n} X_i, \mu \equiv E(X)$ et $0 \leq \delta < 1$, alors :*

$$P(X \geq (1+\delta)\mu) \leq e^{-\frac{\mu\delta^2}{3}} \text{ et } P(X \leq (1-\delta)\mu) \leq e^{-\frac{\mu\delta^2}{2}}$$

3. Etude du patron « invocation parallèle et nombre constant de services élémentaires »

3.1. *Spécification du problème*

Les systèmes distribués et parallèles faiblement couplés sont souvent composés par des modules dont les interactions sont plutôt rares. De plus, les interactions se font par des mécanismes de synchronisation et d'échange de messages. C'est le cas par exemple des services Web composites, basés généralement sur l'information stockée dans des bases de données réparties, offerte par différents fournisseurs de services. Pour satisfaire une requête, un intergiciel peut invoquer les différents fournisseurs et intégrer leur réponse afin de faire une synthèse (Menascé, 1995). Dans le contexte des services Web, ce mécanisme est appelé « fork and merge » (Menascé, 1995). Notons que cela n'est pas un « fork-join » traditionnel, dans le sens où dans notre cas les différentes réponses renvoyées par les services élémentaires font ensuite l'objet d'un traitement par l'intergiciel.

Nous supposons, dans la suite de ce papier, que la distribution du temps de réponse des services élémentaires, participant à la composition, est exponentielle de paramètre λ et que la distribution du temps de réponse du composant de fédération est aussi exponentielle de paramètre μ. Soit n le nombre maximum de services élémentaires participant à la composition. Nous distinguons, dans ce qui suit, deux cas de services Web composites, à savoir :

– Le cas où le composant de fédération attend toutes les réponses des différents services élémentaires pour commencer son traitement.

– Le cas où le composant de fédération traite les réponses au fur et à mesure de leur arrivée.

Dans le premier cas, le temps moyen de réponse du service composite est donné par l'équation suivante :

$$E(T^n_{fixe}) = E(T^n_{syn}) + n\mu \quad [1]$$

où le temps moyen de synchronisation $E(T^n_{syn})$ est donné par le lemme 1.

Lemme 1 *Dans le cas de services élémentaires homogènes de lois de service exponentielles de paramètres $\lambda_i = \lambda$, pour $i \in \{1...n\}$, le temps de synchronisation est donné par :*

$$E(T^n_{syn}) = \frac{1}{\lambda} \sum_{i=1}^{n} \frac{1}{i}$$

Preuve
Le temps moyen pour qu'un premier service achève son exécution est égal à $\frac{1}{n\lambda}$, le temps moyen pour que deux services achèvent leur exécution est égal à $\frac{1}{n\lambda} + \frac{1}{(n-1)\lambda}$ et ainsi de suite. En itérant, on obtient :

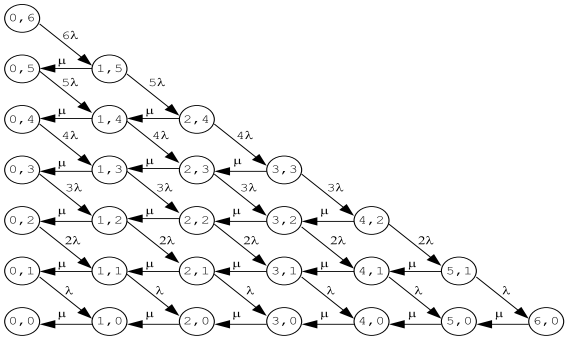
$$E(T^n_{syn}) = \frac{1}{n\lambda} + \frac{1}{(n-1)\lambda} + ... + \frac{1}{2\lambda} + \frac{1}{\lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} \frac{1}{i}$$

c.q.f.d. ◊◊◊

Dans le cas où le composant de fédération traite les réponses au fur et à mesure de leur arrivée, le système peut être décrit par une chaîne de Markov à temps continu $\mathcal{M}$ dont un état $(s_{i,j})$ est défini par :

– i, le nombre de réponses en attente dans la file d'attente modélisant le composant de fédération ;
– j, le nombre de services élémentaires n'ayant pas encore terminé leur service.

Le graphe de transition d'une telle chaîne de Markov, pour le cas où le nombre de services élémentaires participant à la composition est $n = 6$, est donné ci-dessous.



Le calcul du temps moyen de réponse est simplement le temps moyen d'atteinte de l'état $s_{0,0}$ partant de l'état $s_{0,n}$. Plus généralement, pour un état $s_{i,j}$, le temps moyen

d'atteinte de l'état $s_{0,0}$, noté $T(s_{i,j})$, est donné par :

$$T(s_{i,j}) = Soj(s_{i,j}) + \sum_{s_{i,j} \longrightarrow s_{i',j'}} P[s_{i,j}, s_{i',j'}]T(s_{i',j'})$$

avec $Soj(i,j)$ le temps de séjour dans l'état $s_{i,j}$ et $P[s_{i,j}, s_{i',j'}]$ est la probabilité de transition de l'état $s_{i,j}$ à l'état $s_{i',j'}$ déduite du générateur infinitésimal Q.
Ainsi, le temps moyen de réponse d'un tel système peut se calculer en un temps $O(n^2)$ (i.e. le nombre d'équations) et en espace en $O(n)$ (car les $T(s_{\cdot,j})$ sont suffisants pour calculer les $T(s_{\cdot,j+1})$).

3.2. *Modèles génériques de bornes*

Nous proposons dans cette section des modèles bornant le temps moyen de réponse du modèle initial. Ces modèles sont obtenus en « re-dirigeant » certains arcs du modèle initial. Soit $\mathcal{M}$ la chaîne de Markov correspondant au modèle initial. Nous « redirigeons » un certain nombre d'arcs de la chaîne $\mathcal{M}$. Soit $\mathcal{A}$ cet ensemble d'arcs redirigés et $\mathcal{M}_{\mathcal{A}}$ le nouveau processus markovien ainsi obtenu. Pour alléger les notations, nous utilisons dans la suite de ce papier $\mathcal{M}'$ au lieu de $\mathcal{M}_{\mathcal{A}}$. Nous distinguons deux types de redirections, à savoir :

– une redirection des arcs $(s_{i,j}, s_{i+1,j-1})$ de taux $j\lambda$ qui deviennent $(s_{i,j}, s_{i,j})$,
– une redirection des arcs $(s_{i,j}, s_{i-1,j})$ de taux μ qui deviennent $(s_{i,j}, s_{i,j})$,

Nous formalisons ci-dessous ce type de redirection.

Définition 2 *Soit $\mathcal{A}$ un ensemble d'arcs fixé, alors :*
La fonction $suiv_\lambda : S \mapsto S$ est définie par :

$$\forall s_{i,j} \in S, suiv_\lambda(s_{i,j}) = \begin{cases} s_{i,j}, & \textit{si } (s_{i,j}, s_{i+\mathbb{1}_{\{j>0\}},j-\mathbb{1}_{\{j>0\}}}) \in \mathcal{A} \\ s_{i+\mathbb{1}_{\{j>0\}},j-\mathbb{1}_{\{j>0\}}}, & \textit{sinon} \end{cases}$$

La fonction $suiv_\mu : S \mapsto S$ est définie par :

$$\forall s_{i,j} \in S, suiv_\mu(s_{i,j}) = \begin{cases} (s_i, s_j), & \textit{si } (s_{i,j}, s_{i-\mathbb{1}_{\{i>0\}},j}) \in \mathcal{A} \\ s_{i-\mathbb{1}_{\{i>0\}},j}, & \textit{sinon} \end{cases}$$

Le théorème 1 est à la base de la définition de nos familles de modèles bornants.

Théorème 1 *Soient $\mathcal{M}$ la chaîne associée au modèle de « fork and merge » et $\mathcal{M}'$ une chaîne obtenue par redirection d'un ensemble d'arcs $\mathcal{A}$. Alors il existe un couplage $\mathcal{M}^*$ de $\mathcal{M}$ et $\mathcal{M}'$ tel que l'ensemble des états S^* soit défini par :*

$$S^* = \{(s_{i,j}, s_{i',j'}) \mid j \leq j' \wedge i \leq i' + (j' - j)\} \quad [2]$$

Preuve
On définit le couplage à partir de $(s_{i,j}, s_{i',j'}) \in S^*$, en distinguant quatre cas :

1. $i = i'$ et $j = j'$. Dans ce cas, on couple l'état $(s_{i,j}, s_{i,j})$ comme suit :
Si $i > 0$, alors on le couple avec l'état $(s_{i-1,j}, suiv_\mu(s_{i,j}))$, avec un taux μ.
Si $j > 0$, alors on le couple avec l'état $(s_{i+1,j-1}, suiv_\lambda(s_{i,j}))$ avec un taux $j\lambda$.
On le couple avec l'état $(s_{i,j}, s_{i,j})$ avec un taux $(n-j)\lambda + \mathbb{1}_{\{i=0\}}\mu$.

2. $j = j'$ et $i < i'$. Alors $i' > 0$. Dans ce cas, on couple l'état $(s_{i,j}, s_{i',j'})$ comme suit :
Si $i > 0$, alors on le couple avec l'état $(s_{i-1,j}, suiv_\mu(s_{i',j}))$ avec un taux μ, sinon on le couple avec l'état $(s_{i,j}, suiv_\mu(s_{i',j}))$ avec un taux μ.
Si $j > 0$, alors on le couple avec l'état $(s_{i+1,j-1}, suiv_\lambda(s_{i',j}))$ avec un taux $j\lambda$.
On le couple avec l'état $(s_{i,j}, s_{i',j})$ avec un taux $(n-j)\lambda + \mathbb{1}_{\{i=0\}}\mu$.

3. $j < j'$ et $i + j = i' + j'$. Alors $i > 0$ et $j' > 0$. Dans ce cas, on couple l'état $(s_{i,j}, s_{i',j'})$ comme suit :
On le couple avec l'état $(s_{i-1,j}, suiv_\mu(s_{i',j'}))$, avec un taux μ.
Si $j = 0$, alors on le couple avec l'état $(s_{i,j}, suiv_\lambda(s_{i',j'}))$ avec un taux $j'\lambda$.
Si $j > 0$, alors on le couple avec :

- l'état $(s_{i+1,j-1}, suiv_\mu(s_{i',j'}))$, avec un taux $j\lambda$,
- et l'état $(s_{i,j}, suiv_\mu(s_{i',j'}))$, avec un taux $(j'-j)\lambda$.

On le couple avec l'état $(s_{i,j}, s_{i',j'})$ avec un taux $(n-j)\lambda + \mathbb{1}_{\{i=0\}}\mu$.

4. $j < j'$ et $i + j < i' + j'$. Alors $i' > 0$ et $j' > 0$. Dans ce cas, on couple l'état $(s_{i,j}, s_{i',j'})$ comme suit :
Si $i > 0$, alors on le couple avec l'état $(s_{i-1,j}, suiv_\mu(s_{i',j'}))$ avec un taux μ, sinon on le couple avec l'état $(s_{i,j}, suiv_\mu(s_{i',j'}))$ avec un taux μ.
Si $j = 0$, alors on le couple avec l'état $(s_{i,j}, suiv_\lambda(s_{i',j'}))$ avec un taux $j'\lambda$.
Si $j > 0$, alors on le couple avec :

- l'état $(s_{i+1,j-1}, suiv_\mu(s_{i',j'}))$, avec un taux $j\lambda$,
- et l'état $(s_{i,j}, suiv_\mu(s_{i',j'}))$, avec un taux $(j'-j)\lambda$.

On le couple avec l'état $(s_{i,j}, s_{i',j'})$ avec un taux $(n-j)\lambda + \mathbb{1}_{\{i=0\}}\mu$.

Le fait que $\mathcal{M}^*$ est le couplage recherché se fait par un examen des taux marginaux dans chacun des cas.

c.q.f.d. $\Diamond\Diamond\Diamond$

Le corollaire suivant est une conséquence immédiate de la proposition 1.

Corollaire 1 *Soit* $(s_{i,j}, s_{i',j'}) \in S^*$*, alors :*

$$T(s_{i,j}) \leq T'(s_{i',j'})$$

Et en particulier, $T(s_{0,n}) \leq T^{'}(s_{0,n})$.

Preuve
Il suffit de remarquer que $\forall (s, s') \in S^*\ s' = s_{0,0} \Rightarrow s = s_{0,0}$ et d'appliquer la proposition 1.

c.q.f.d. ◊◊◊

Soient $\mathcal{A}_1 \subseteq \mathcal{A}_2$ deux ensembles d'arcs redirigés et $\mathcal{M}_{\mathcal{A}_1}$, $\mathcal{M}_{\mathcal{A}_2}$ les deux chaînes résultantes. Le théorème 2 qui a une preuve similaire à celle du théorème 1 montre que la redirection est « croissante » suivant l'ordre (partiel) de l'inclusion.

Théorème 2 *Soient $\mathcal{M}$ la chaîne associée au modèle de « fork and merge » et $\mathcal{M}_{\mathcal{A}_1}$ (resp. $\mathcal{M}_{\mathcal{A}_2}$) une chaîne obtenue par redirection d'un ensemble d'arcs $\mathcal{A}_1$ (resp. $\mathcal{A}_2$) avec $\mathcal{A}_1 \subseteq \mathcal{A}_2$. Alors il existe un couplage $\mathcal{M}^*$ de $\mathcal{M}_{\mathcal{A}_1}$ et $\mathcal{M}_{\mathcal{A}_2}$ tel que l'ensemble des états S^* soit défini par :*

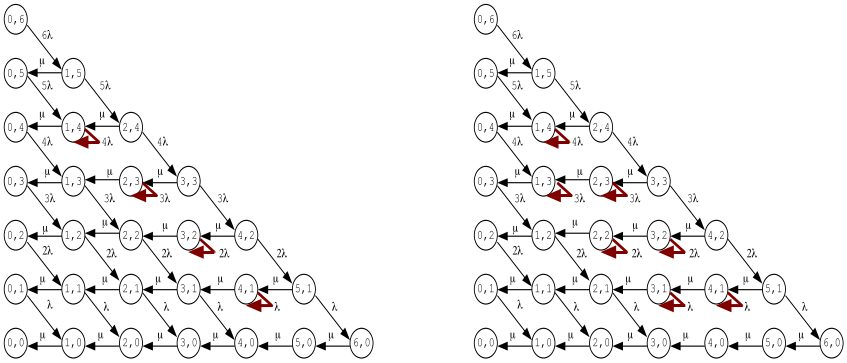
$$S^* = \{(s_{i,j}, s_{i',j'}) \mid j \leq j' \wedge i \leq i' + (j' - j)\} \quad [3]$$

3.3. *Familles particulières de modèles de bornes*

Nous proposons ici trois familles de modèles bornants obtenues par un choix spécifique d'ensemble d'arcs redirigés. Il peut sembler surprenant que l'on obtienne une réduction de la complexité de résolution alors que l'espace d'états des chaînes bornantes est le même que l'espace d'états initial. Le point clef est que la structure particulière des chaînes bornantes permet de « fusionner » plusieurs équations en une seule selon le paramètre d'agrégation lié à chaque famille de modèles bornants. Par manque de place, nous ne détaillons pas ces équations.

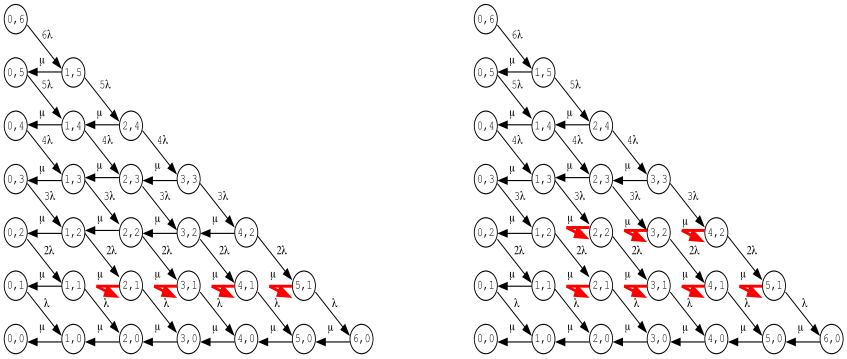
3.3.1. *Bornes par λ-bouclage*

Ces modèles de bornes sont caractérisés par un paramètre m pour $0 \leq m \leq n-2$ que l'on appelle niveau d'agrégation. Il consiste à « stopper » les services élémentaires dès qu'au moins une requête a été traitée et jusqu'à ce que : soit il n'y ait plus de requêtes à traiter, soit $m+1$ requêtes aient été traitées. Nous avons présenté ci-dessous, deux modèles bornants par λ-bouclage pour $n = 6$ et pour $m \in \{1, 2\}$.



3.3.2. *Bornes par μ-bouclage final*

Ces modèles de bornes sont caractérisés par un paramètre m pour $0 \leq m \leq n-2$ que l'on appelle niveau d'agrégation. Il consiste à « stopper » la fusion de services dès qu'il reste au plus m requêtes en cours et jusqu'à ce qu'il n'y ait plus de requêtes en cours. Nous avons présenté ci-dessous, deux modèles bornants par μ-bouclage final pour $n = 6$ et pour $m \in \{1, 2\}$.



3.3.3. *Bornes par μ-bouclage initial*

Ces modèles de bornes sont caractérisés par un paramètre h que l'on appelle niveau d'agrégation. Il consiste à « stopper » la fusion de services tant qu'il reste plus de h requêtes en cours.

4. Résultats numériques pour le « fork-merge » fixe

Nous présentons dans cette section un récapitulatif des résultats numériques obtenus pour les familles présentées ci-dessus. Il est clair que si $\mu \ll \lambda$ (resp. $\mu \gg n\lambda$) les bornes par λ-bouclage sont meilleures que les bornes par μ-bouclage et vice versa. Ainsi, le cas le plus intéressant et que nous étudions dans ce qui suit est celui où $\lambda \leq \mu \leq n\lambda$. Ce qui détermine le comportement quantitatif de la chaîne de Markov étudiée étant le rapport entre μ et $n\lambda$, nous fixons alors le paramètre λ à 0,1 et faisons varier la valeur du taux de fédération μ, sans perte de généralité.

4.1. *Comparaison de λ-bouclage et μ-bouclage final*

Nous calculons, dans cette section, les bornes proposées pour plusieurs services Web composites dont le nombre de services élémentaires participant à sa composition n est égal à 5.000, 10.000, 15.000 et 20.000. Le niveau d'agrégation $m = O(n - \sqrt{n})$. Ainsi, le nombre d'équations de récurrence nécessaires pour calculer le temps d'atteinte de l'état $s_{0,0}$ dans chacun des modèles considérés est en $O(n\sqrt{(n)})$. Les

figures 1 et 2 montrent l'évolution du temps moyen de réponse exact d'un service Web composite et les bornes supérieures obtenues dans le cas de λ-bouclage et μ-bouclage, en fonction du taux de fédération μ-bouclage, pour différentes valeurs du nombre de services élémentaires participant à sa composition. Ainsi, d'après ces résultats nous pouvons conclure que :

– la famille de modèles λ-bouclage est meilleure que la famille de modèles μ-bouclage,

– la famille de modèles λ-bouclage donne de meilleurs résultats lorsque le taux de fédération μ n'est pas très grand par rapport au taux de service λ.

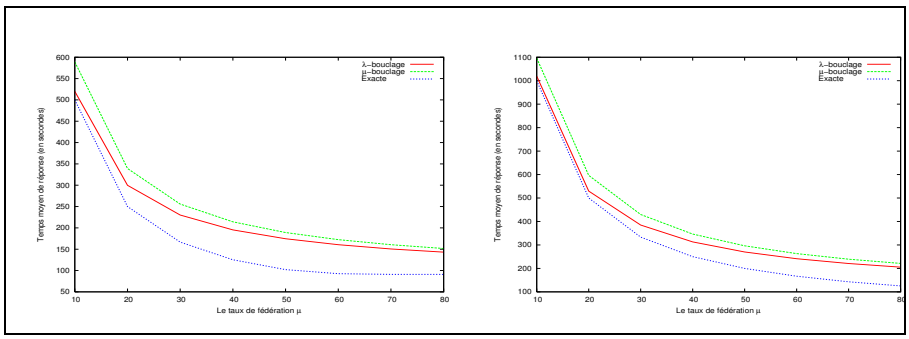


Figure 1. *Nombre de services* $n = 5.000$ *(à gauche) et* $n = 10.000$ *(à droite)*

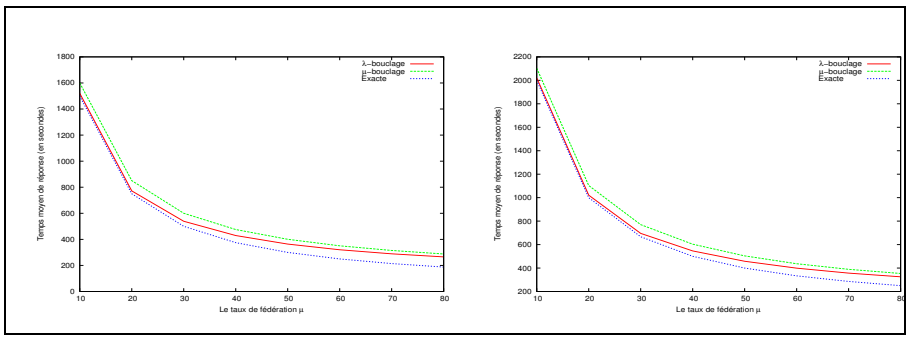


Figure 2. *Nombre de services* $n = 15.000$ *(à gauche) et* $n = 20.000$ *(à droite)*

4.2. *Comparaison du λ-bouclage et μ-bouclage initial*

Nous calculons les bornes proposées (λ-bouclage et μ-bouclage initial) pour plusieurs services Web composites dont le nombre de services élémentaires participant à sa composition n égal à 5.000, 10.000, 15.000 et 20.000. Ainsi, d'après les résultats obtenus (voir figures 3, 4 nous observons que la borne par μ-bouclage initial est meilleure que λ-bouclage quand le taux de fédération est très grand par rapport au taux de service.

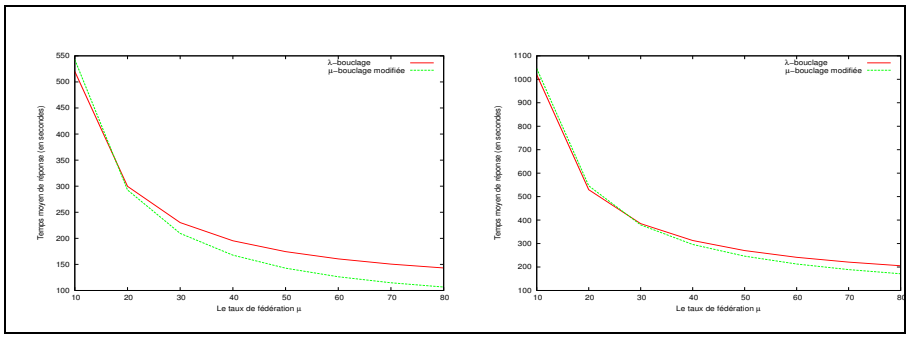


Figure 3. *Nombre de services n=5.000 (à gauche) et n=10.000 (à droite)*

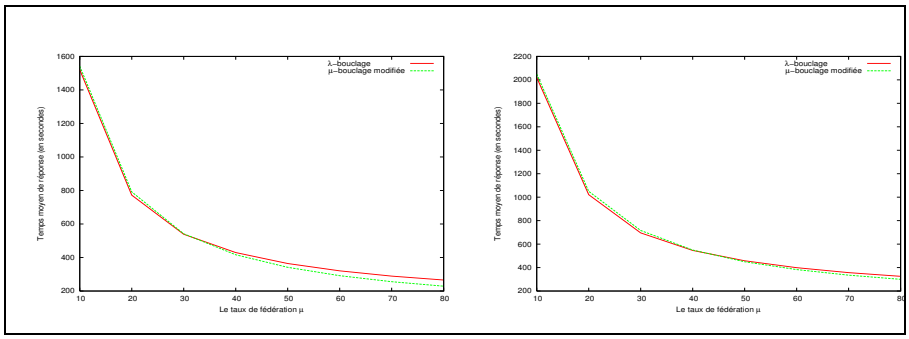


Figure 4. *Nombre de services n=15.000 (à gauche) et n=20.000 (à droite)*

4.3. *Compromis entre la qualité de la borne et la complexité de calcul*

Nous montrons dans ce qui suit le compromis entre la qualité de la borne et la complexité de calcul dans le cas où la borne par μ-bouclage initial est meilleure que λ-bouclage. Le tableau 1 montre le compromis entre la qualité de la borne et la complexité de calcul, avec l'erreur relative et le ratio définit comme étant le nombre d'équations de récurrence lorsque l'on considère la chaîne de Markov initiale $\mathcal{M}$ et la chaîne de Markov modifiée $\mathcal{M}'$. Le nombre de services élémentaires $n = 20.000$, $\lambda = 0.1$ et $\mu = 70$ (μ-bouclage modifiée est meilleure que λ-bouclage). La valeur exacte du temps moyen de réponse de tel service Web composite est $E(T^n_{fixe}) = 285.704$. Le nombre d'équations de récurrence nécessaires pour le calcul du temps de réponse exacte est 200030001 équations. Ainsi, plus le niveau d'agrégation h augmente, plus le nombre d'équations de récurrence nécessaires pour calculer le temps d'atteinte de l'état $s_{0,0}$ partant de l'état $s_{0,n}$ augmente et plus la qualité de la borne sera meilleure.

Niveau d'agrégation (h)	Ratio	Valeur de la borne	Erreur relative
0	0,02%	390,47	36,84%
100	1,01%	338,62	18,52%
500	4,95%	322,58	12,90%
1000	9,76%	315,66	10,49%
1200	11,65%	313,83	9,85%
1400	13,52 %	312,29	9,30%
1600	15,38 %	310,96	8,84%
1800	17,21 %	309,78	8,43%

Tableau 1. *Compromis entre la qualité de la borne et la complexité de calcul*

5. Etude du patron « invocation parallèle et nombre aléatoire de services élémentaires »

5.1. *Spécification du problème*

Nous considérons dans cette section le cas où chaque service élémentaire a la même probabilité d'être invoqué et ceci de manière indépendante des autres services. Nous distinguons deux cas, à savoir :

– Le cas où le composant de fédération attend toutes les réponses des différents services Web élémentaires invoqués pour commencer le traitement des différentes réponses des services élémentaires.

– Le cas où le composant de fédération traite les réponses des services Web élémentaires participant à la composition au fur et à mesure de leurs arrivées.

Dans le premier cas, le temps de réponse moyen du service composite est donné par l'équation suivante (Haddad *et al.*, 2008) :

$$E(T_{var}^{n,p}) = E(T_{var}^{n,p}(syn)) + \mu \sum_{i=1}^{n} P(X = i) = \sum_{i=1}^{n} P(X = i) \sum_{j=1}^{i} \frac{\lambda}{i} \qquad [4]$$

Dans le second cas, nous combinons la technique des familles de modèles de bornes proposées pour le cas où le nombre de services Web élémentaires est constant avec une majoration obtenue par application des bornes de Chernoff.

5.2. *Modèles de bornes*

Nous notons $p < 1$, la probabilité d'invocation d'un service Web élémentaire. Observons tout d'abord que le temps de réponse moyen d'un service Web composite augmente lorsque le nombre de services élémentaires qui le composent augmente. Cette observation intuitive se démontre facilement à l'aide du corollaire 1 appliqué avec $\mathcal{M}' = \mathcal{M}$ où $\mathcal{M}$ est la chaîne correspondant au plus grand nombre de services et en remarquant que l'état initial de la chaîne correspondant au plus petit nombre de services est couplé avec l'état initial de $\mathcal{M}$. La proposition suivante établit alors la borne recherchée.

Proposition 3 *Soit X, le nombre de services élémentaires invoqués, variable aléatoire binomiale de paramètres (n, p). Soit $np < n' \leq n$ et posons $\delta = \frac{n'-np}{np}$. Alors :*

$$E(T_{var}^{n,p}) \leq E(T_{fixe}^{n'}) + e^{-\frac{\delta^2 np}{3}} E(T_{fixe}^{n}) \quad [5]$$

Preuve
Le théorème des probabilités totales permet d'écrire :

$$E(T_{var}^{n,p}) = E(T_{var}^{n,p} \mid X \leq n')P(X \leq n') + E(T_{var}^{n,p} \mid X > n')P(X > n')$$

En appliquant deux fois la borne dans le cas d'un nombre de services fixe, en majorant $P(X \leq n')$ par 1 et puis en appliquant la borne de Chernoff, on obtient :

$$E(T_{var}^{n,p}) \leq E(T_{fixe}^{n'}) + e^{-\frac{\delta^2 np}{3}} E(T_{fixe}^{n})$$

c.q.f.d. ◊◊◊

Nombre de services élémentaires	TE	Exacte	Borne	Erreur relative	δ-optimale
5.000	1087,69	74,86	76,08	1,63%	0,12
8.000	4444,00	79,56	80,57	1,27%	0,10
10.000	8709,96	81,79	82,80	1,23%	0,09
15.000	29291,60	85,89	88,653	3,21%	0,07

Tableau 2. *Temps de calcul de la valeur exacte du temps moyen de réponse, de la valeur de la borne et sa qualité dans le cas où $p = 0,2$*

6. Résultats numériques pour le « fork-merge » variable

Nous présentons, dans cette section, les résultats numériques obtenus dans le cas où le nombre de services Web élémentaires, participant à la composition d'un service, est une variable aléatoire. Les courbes de la figure 5 montrent l'évolution de la valeur optimale de δ en fonction de la probabilité d'invocation d'un service Web élémentaire. Pour obtenir la valeur optimale de δ, nous avons calculé pour une probabilité d'invocation donnée p (fixe) la meilleure borne et puis nous avons déduit la meilleure valeur de la borne (en faisons varier δ). D'après les résultats obtenus, nous pouvons remarquer que :

Nombre de services élémentaires	TE	Valeur exacte	Borne	Erreur relative	δ-optimale
5.000	1087,69	88,73	97,66	10%	0,06
8.000	4444,00	93,84	127,74	36,12%	0,04
10.000	8709,96	103,35	149,16	44,32%	0,04
15.000	29291,60	150,058	201,89	34,54%	0,03

Tableau 3. *Temps de calcul de la valeur exacte du temps moyen de réponse, de la valeur de la borne et sa qualité dans le cas où $p = 0,8$*

– lorsque la probabilité d'invocation est très petite, δ optimale est très proche de 1,

– la valeur optimale de δ décroît avec l'augmentation du nombre de services élémentaires invoqués.

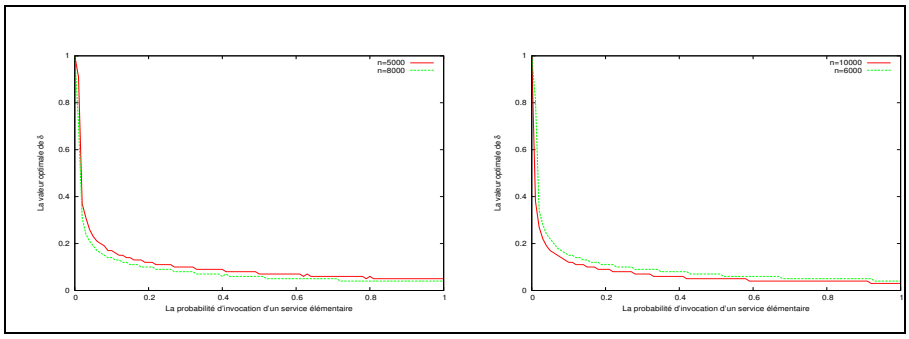


Figure 5. *Evolution de la valeur optimale de δ en fonction de la probabilité d'invocation, dans les cas $n = 5000, n = 8000$ (à gauche) et $n = 6000, n = 1000$ (à droite)*

Un récapitulatif, qui donne une indication sur la qualité de la borne et le gain considérable sur le temps d'exécution, est fourni par les tableaux 2 et 3 (notons que le calcul de la borne est instantané) dans les cas respectivement de probabilité égale à 0,2 et 0,8. Les valeurs des taux sont fixées à $\lambda = 0,1$ et $\mu = 80$. Ainsi, les bornes par μ-bouclage initial sont meilleures que les bornes par λ-bouclage. Dans ces tableaux, TE représente le temps d'exécution en secondes pour le calcul de la valeur exacte du temps moyen de réponse d'un service Web composite, Valeur exacte la valeur exacte du temps moyen de réponse du service Web composite, Erreur relative l'erreur relative entre la valeur exacte du temps moyen de réponse et la borne obtenue et δ-optimale la valeur optimale du paramètre δ. La valeur optimale du paramètre δ est obtenue comme suit : pour une probabilité d'invocation p donnée, d'un service Web élémentaire quelconque, on fait varier le paramètre δ par pas de 0,001 et on retenir la meilleure borne obtenue (i.e. la plus petite borne).

7. Conclusion et travaux futurs

Nous avons présenté dans ce papier une approche basée sur le couplage des processus stochastiques pour calculer des bornes supérieures du temps de réponse d'un service Web composite. L'intérêt de notre approche provient de la réduction des temps de calcul. De plus, celle-ci permet de trouver un compromis entre la qualité de la borne et le temps nécessaire à son calcul numérique. Nous avons proposé trois familles de modèles bornant supérieurement le temps de réponse d'un service Web composite. Nous avons aussi proposé la combinaison de ces familles de bornes avec les bornes de Chernoff afin de prendre en compte un nombre aléatoire de services élémentaires

participant à une composition. Par notre méthode dans ce dernier cas, nous nous ramenons à deux cas d'invocation seulement.

Nous envisageons deux extensions à ce travail. Premièrement, généraliser l'étude au cas des patrons plus élaborés (prendre en compte le fait que par exemple les services participant à une composition peuvent être eux-mêmes des services composites). Deuxièmement, prendre en compte le fait que les services Web élémentaires participant à la composition sont hétérogènes. Dans ce dernier cas, nos méthodes pourraient décroître de manière exponentielle la complexité de résolution.

8. Bibliographie

Bray T., Paoli J., Sperberg-McQueen C., Maler E., Yergeau F., « Extensible Markup Language (XML) 1.0 », 2004.

Bussler C., Fensel D., Maedche A., « A conceptual architecture for semantic web enabled web services », *SIGMOD Rec.*, vol. 31, n° 4, p. 24-29, 2002.

Curbera F., Silva-Lepe I., Weerawarana S., « On the integration of heterogeneous web service partners », *In International Semantic Web Conference*, 2001.

Doisy M., Comparaison de processus markoviens, Thèse de doctorat, Université de Pau et des Pays de l'Adour, 1992.

Haddad S., Mokdad L., Youcef S., « Response time analysis of composite Web services », *CSNDSP 2008, Communication Systems, Networks and Digital Signal Processing, IEEE Computer Society, Graz University of Technology, Austria*, July, 2008.

Massy W., « Stochastic ordering for markov processes on partially ordered space », *Mathematics of operation research*, vol. 12, p. 350-367, 1986.

Menascé D., « Static and Dynamic processor Scheduling Discipline in Heterogenous Parallel Architecture », *Parallel and Distributed Computing*, 1995.

Menascé D., « QoS Issues in Web Services », *IEEE Internet Computing*, 2002.

Menascé D., Dubeye V., « Utility-based QoS brokering in service oriented architectures », 2007.

Menascé D., Mason G., « Response-Time Analysis of Composite Web Services », *IEEE Internet computing*, 2004.

Michael S., « On the response Time of Large-scale Composite Web Services », *IEEE*, 2005.

Stoyan D., « Comparison methods for queue and other stochastics models », *J-Wiley and Son*, 1976.

Yang J., Papazoglou M., « Web components : A substrate for web service reuse and composition », *In Proceedings of the 14th International Conference on Advanced Information Systems Engineering, Toronto, Canada*, 2001.