

Le calcul de la disponibilité des services (SLA)

Martin BECH

Danish e-Infrastructure Cooperation (DeiC)
DTU, Bldg. 305
2800 Lyngby
DANEMARK

Résumé

Dans cette contribution, on apprend comment calculer la disponibilité garantie (rapport entre le temps de fonctionnement d'un service et le temps total) d'un service composé, c'est-à-dire un service qui est le produit d'une combinaison de plusieurs services qui ont chacun leur propre disponibilité.

Si l'on a, par exemple, deux lignes WAN qui ont des disponibilités garanties de 99,7% et 99,5% respectivement, on voit souvent ces nombres traités comme des probabilités, et pour une connexion en série, la disponibilité sera calculée comme le produit $99,7\% \cdot 99,5\%$. Ce résultat n'est pas tout à fait exact, et la présente contribution vise à montrer comment calculer des disponibilités garanties de manière plus précise.

Ces dernières années, nos utilisateurs nous demandent de remplacer les services fournis sur la base du « best effort » par des services avec des garanties fermes au niveau de la disponibilité comme un élément d'un SLA.

Le but de cette intervention est de donner une vue générale des règles de calcul de la disponibilité qui permettent d'estimer la disponibilité que l'on peut promettre pour un service qui dépend d'autres services. Il est aussi expliqué comment calculer les disponibilités que l'on doit demander à ses fournisseurs si l'on doit atteindre un niveau donné de disponibilité pour un service composé. Enfin, la relation entre la disponibilité garantie et le temps moyen entre pannes (MTBF) est expliquée.

Ces règles – et d'autres – du « calcul de disponibilité » sont déduites sur la base d'observations provenant du réseau de la recherche danois. Les méthodes statistiques utilisées ne sont pas extraordinaires, mais sur cette discipline spécifique, la littérature (même sur l'internet) apparaît assez limitée.

Mots-clefs

SLA (Service Level Agreement), disponibilité, temps moyen entre pannes (MTBF), services composés, gestion des contrats.

1 Introduction

Une vague de « professionnalisme » déferle sur la gestion de l'informatique ces dernières années – même dans notre secteur universitaire. Désormais, l'offre aux utilisateurs de services « best effort » n'est plus suffisante – il faut garantir une disponibilité minimum.

Si l'on achète un service (comme une ligne WAN ou l'hébergement d'un site web) à un fournisseur qui garantit une certaine disponibilité, et que ce service est offert à nouveau à ses propres utilisateurs, il est simple de promettre le même niveau de disponibilité.

Mais dans les cas où la disponibilité d'un service dépend de celle de plusieurs autres services, il semble que beaucoup de professionnels de l'informatique n'effectuent pas avec rigueur le calcul de la disponibilité et les niveaux de confiance associés.

Afin de remédier à cette situation, les règles du calcul de la disponibilité sont présentées dans la présente contribution, en s'appuyant sur des exemples et des observations qui viennent du réseau de la recherche danois (DeiC).

2 Présentation générale

Dans les contrats que nous avons avec nos fournisseurs, nous disposons de disponibilités garanties. Typiquement, les SLA disent que la disponibilité garantie est de 99,7% ou 99,9%, mesurée sur un trimestre. Voyez ces exemples :

3. Service Assurance

3.1 Service Availability

Service Availability is guaranteed at 99.9%. Spin will provide a service fee rebate to customers with unavailability of greater than 40 minutes in a given month.

“Service Availability” is defined as the percentage of time the service is available, via the primary connectivity medium, during the course of a month. The Service Availability is calculated in accordance with the following formula:

$$\text{Service Availability} = \frac{\text{Total minutes for the period minus Unavailable minutes}^*}{\text{Total minutes for the period}} \times 100$$

* Unavailable minutes is the total number of minutes that the service is unavailable due to issues with the Spin network or our carrier's except for programmed outages.

The following table highlights at-a glance the main WAN and Internet services and their corresponding Service level parameters:

SERVICES	SERVICE LEVEL PARAMETER	NETWORK AVAILABILITY
Ethernet Fibre WAN (E10, E100, FastE, etc)	7 x 24 - SLA	99.7%

2.1 99.7% Service Availability - Subject to Section 5 below, if the availability of the Services is less than 99.7%, Red Fox Hosting will issue a credit to customer in accordance with the following schedule, with the credit being calculated on the basis of the monthly service charge for the affected Services:

Service Availability	Credit Percentage
99.7 to 100%	0%
98% to 99.7%	10%
95% to 97.9%	20%
90% to 94.9%	30%
89.9% or below	100%

Ici, les fournisseurs promettent une disponibilité de 99,7%. Toutefois, cela ne signifie pas que l'on peut s'attendre à ce que le service soit en panne 0,3% du temps en moyenne – ce chiffre correspond simplement à la limite à partir de laquelle le fournisseur doit payer des compensations financières.

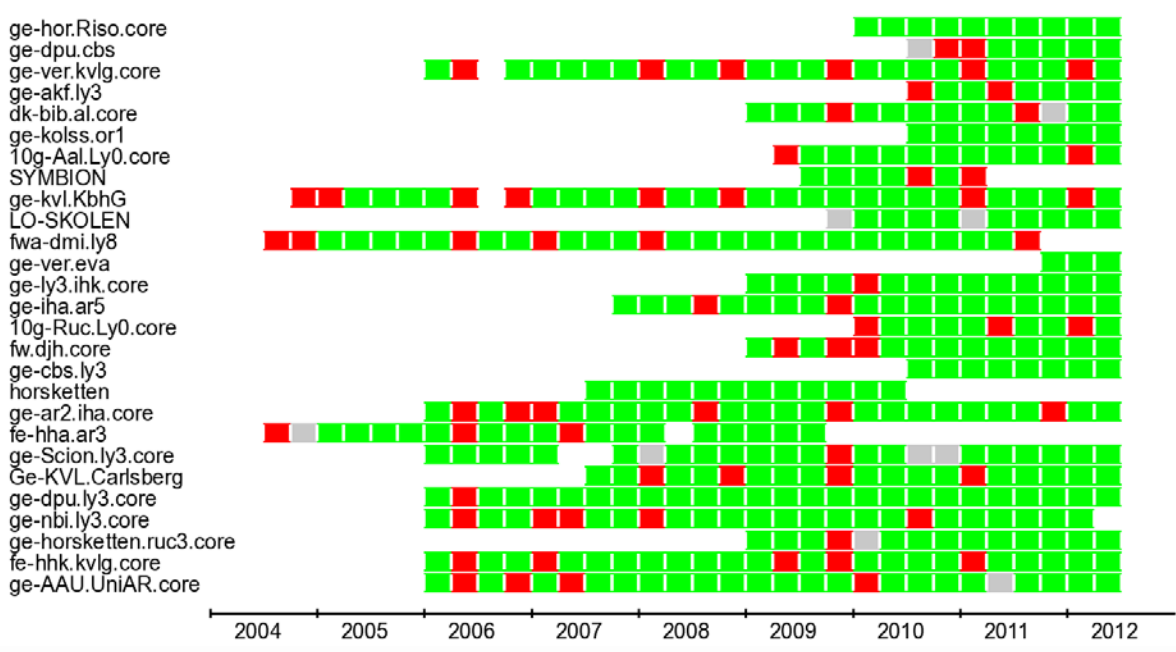
Il est probable que le service soit disponible en moyenne pendant une part plus importante du temps. Pour des simples connexions WAN, nos données montrent que le temps pendant lequel celles-ci sont hors-service est en moyenne d'environ 0,11%. Dans nos contrats, le temps pendant lequel les services ne fonctionnent pas est calculé sur une base trimestrielle et cela correspond au fait que le fournisseur ne respecte pas les niveaux contractuels pendant 6,2% des trimestres, soit un trimestre défaillant sur une période de 4 ans.

Pour chaque connexion WAN, nous mesurons automatiquement toutes les dix minutes si la connexion fonctionne au moyen d'une série de commandes ping. S'il y a une perte de paquets ou si le temps de réponse est trop long, la connexion

est considérée comme en panne. Pour chaque ligne, par trimestre, on dispose donc d'une série de mesures indiquant que la connexion est en panne, dont le nombre est N_{panne} tandis que N_{ordre} désigne le nombre de mesures indiquant que la connexion fonctionne. Nous calculons la disponibilité :

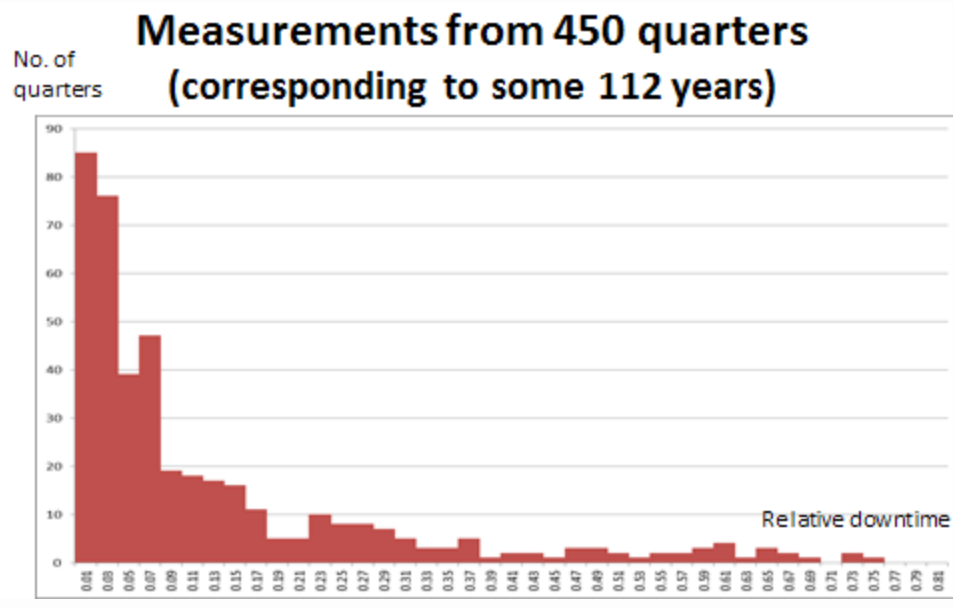
$$\frac{N_{panne}}{N_{panne} + N_{ordre}}$$

Effectué pour chaque ligne, cela donne une représentation du temps de fonctionnement, de la manière suivante :



Ici, les cases rouges représentent les trimestres pendant lesquels la disponibilité était inférieure à 99,7%, les cases vertes les trimestres pendant lesquels la disponibilité était supérieure à ce niveau et les cases grises les trimestres pour lesquels, pour des raisons différentes, nous ne disposons pas de données valides. Les mesures proviennent du réseau danois de la recherche.

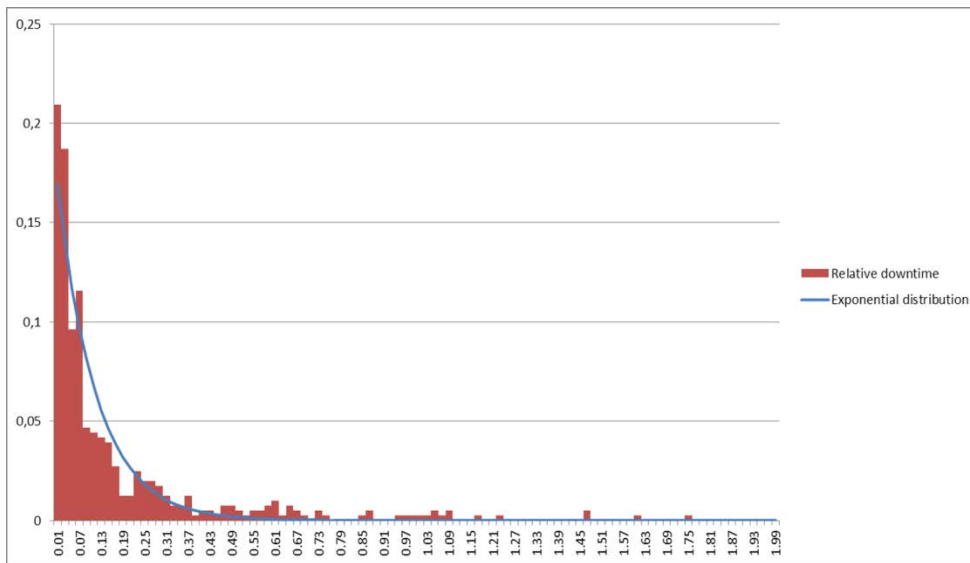
Si l'on classe ces données en fonction du temps de panne relatif, on obtient le diagramme suivant :



Dans ce diagramme, nous avons toutefois exclu certains trimestres ayant connu des temps de panne très importants (exceptions statistiques). Il est raisonnable d'exclure ces mesures parce qu'elles s'expliquent par une série de raisons qui sont hors de contrôle du fournisseur ou qui ne pouvaient avoir été prévues par le fournisseur :

- Des erreurs dans nos mesures peuvent être dues entre autres au fait qu'en pratique, il y a eu un délai entre la mise hors service d'une ligne et l'arrêt de nos mesures
- Des erreurs peuvent être dues à une maintenance prévue pendant laquelle nous avons continué à effectuer des mesures
- Certains évènements de « force majeure » ne peuvent pas être considérés comme prévisibles dans le contrat avec le fournisseur.

Sous la forme d'un diagramme de fréquences, nos données ressemblent à une distribution exponentielle :



Pour une distribution exponentielle, nous avons :

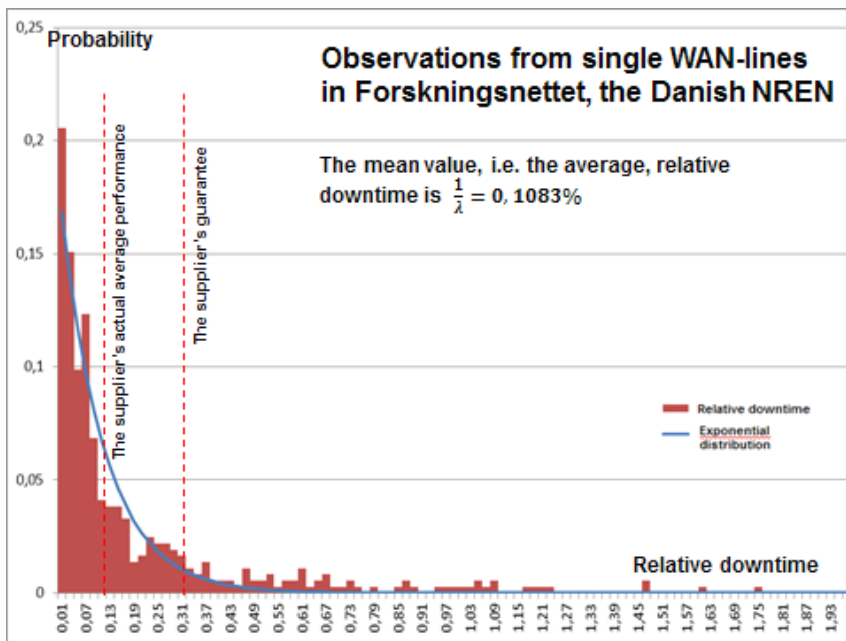
$$\text{Fonction de répartition } F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

$$\text{Fonction de densité } f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

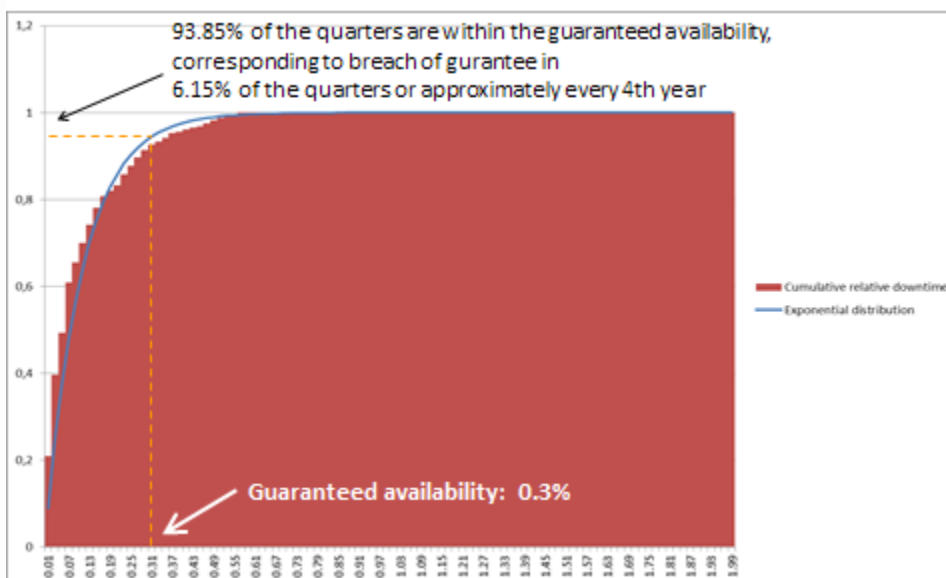
L'espérance ou la moyenne est $E(x) = \frac{1}{\lambda}$ (qui peut aussi être utilisée pour trouver λ)

La distribution exponentielle est « sans mémoire », ce qui veut dire que pour une variable aléatoire X on a $P(X > t + v | X > v) = P(X > t) \Leftrightarrow P(X > t + v) = P(X > t) \cdot P(X > v)$

Pour le dire autrement, la probabilité d'une heure supplémentaire de panne ne dépend pas du fait qu'il y ait eu des pannes auparavant. La distribution exponentielle est la seule distribution statistique continue qui possède cette propriété. Comme l'on peut considérer raisonnablement que les pannes dont on parle ici sont dotées de cette propriété, on aurait pu s'attendre à ce que les données décrites ci-dessus suivent cette loi exponentielle.



Comme nous trouvons le paramètre $\lambda = 0,1075\%$ dans les données, on observe dans la fonction de répartition que le temps de panne maximal garanti de 0,3% correspond à ce que $F(0,3\%) = 93,85\%$ des trimestres aient un temps de panne inférieur à ce niveau. On peut aussi dire que la disponibilité est inférieure au niveau garanti dans $1 - 93,85\% = 6,15\%$ des trimestres. Cela correspond au fait qu'un trimestre sur 16 est défaillant c'est-à-dire non conforme au contrat environ une fois par période de quatre ans. Il est donc raisonnable de définir un paramètre MTBF $\theta = 16$ (trimestres).



Si la disponibilité garantie (ici 99,7%) est G , et que nous supposons que ces données suivent la loi exponentielle, on peut déterminer le temps moyen de fonctionnement (performance) p , de la manière suivante:

$$1 - \frac{1}{\theta} = F(1 - G) = 1 - e^{-\lambda(1-G)} \Leftrightarrow -\ln\theta = -\lambda(1 - G) \Leftrightarrow \lambda = \frac{\ln\theta}{1 - G}$$

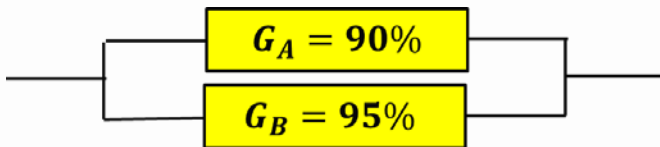
et donc

$$p = 1 - \frac{1}{\lambda} = 1 - \frac{1 - G}{\ln\theta}$$

Soit une relation linéaire entre p et G .

Nous pouvons donc maintenant traduire les disponibilités garanties en probabilités et déterminer la probabilité (ici la disponibilité) pour un service combiné et ainsi calculer la disponibilité garantie pour ce dernier.

Considérez, par exemple, une situation où l'on combine deux lignes WAN en parallèle afin d'obtenir un niveau plus élevé de disponibilité, et que les disponibilités garanties sont de 90% et 95% respectivement :



De fait, ce que l'on doit calculer est la probabilité pour qu'au moins une de ces lignes fonctionne :

$$p_{AVB} = 1 - (1 - p_A) \cdot (1 - p_B)$$

Alors nous convertissons d'abord les temps de fonctionnement garantis en disponibilité réelle:

$$p_A = 1 - \frac{1-G_A}{\ln\theta} = 96,393\% \quad \text{respectivement} \quad p_B = 1 - \frac{1-G_B}{\ln\theta} = 98,197\%$$

On calcule ensuite la disponibilité réelle qui en résulte

$$p_{AVB} = 1 - (1 - p_A) \cdot (1 - p_B) = 99,935\%$$

Et nous la re-convertissons en disponibilité garantie que nous pouvons promettre à l'utilisateur du service composé :

$$G_{AVB} = 1 - (1 - p_{AVB})\ln\theta = 99,8\%$$

Si l'on regardait les disponibilités garanties comme des probabilités, on pourrait calculer la disponibilité comme :

$$G_{AVB} = 1 - (1 - G_A) \cdot (1 - G_B) = 1 - (1 - 90\%) \cdot (1 - 95\%) = 99,5\%$$

Ceci n'est pas tout à fait exact car les disponibilités garanties ne sont pas des probabilités, mais des niveaux de confiance, et l'on voit maintenant que le résultat est plutôt

$$G_{AVB} = 1 - \frac{(1 - G_A)(1 - G_B)}{\ln\theta} = 99,8\%$$

où θ est ce paramètre de temps moyen entre pannes (MTBF), que des données observées dans le réseau de la recherche danois situe à approximativement 16 pour les lignes WAN normales.

De même, la règle pour une combinaison en série :



n'est pas

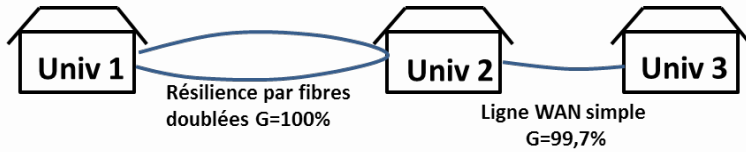
$$G_{A \wedge B} = G_A \cdot G_B$$

mais plutôt

$$G_{A \wedge B} = 1 - \ln\theta \left(1 - \left(1 - \frac{1 - G_A}{\ln\theta} \right) \left(1 - \frac{1 - G_B}{\ln\theta} \right) \right)$$

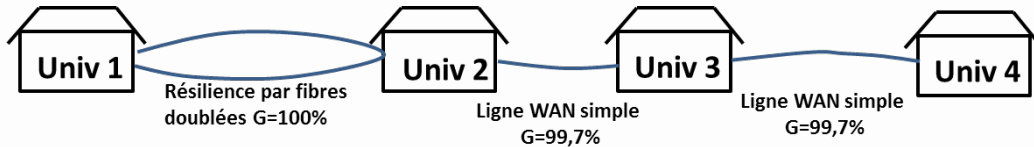
Ces règles peuvent être utilisées dans des situations plus complexes, comme le lecteur pourra le constater par lui-même.

Une autre utilisation de la méthode utilisée ici peut consister à déterminer le MTBF pour une combinaison donnée de services. Si l'on a par exemple une situation comme celle-ci:



la disponibilité garantie pour l'ensemble de la connexion de Univ1 à Univ3 est de 99,7%.

Si l'on prolonge cette connexion avec une autre fibre, la situation est alors la suivante :



Normalement, nous calculerions la disponibilité réelle en résultant comme $p_{total} = (1 - \frac{1-G}{\ln\theta})^2 = 99,78\%$ et la disponibilité garantie comme $G_{total} = 1 - \ln\theta \left(1 - \left(1 - \frac{1-G}{\ln\theta}\right)^2\right) = 99,4\%$.

Que se passerait-il dans cette situation si malgré cela, nous garantissons à Univ4 une disponibilité de 99,7%? Dans cette situation, à quelle fréquence ne respecterions-nous pas nos engagements ? Pour le déterminer, considérons :

$$G = 1 - (1 - p)\ln\theta \Leftrightarrow \theta = e^{\frac{1-G}{1-p}} = e^{\frac{1-99,7\%}{1-99,78\%}} = 4$$

Si nous sommes prêts à être en faute un trimestre sur 4, nous pourrions donc garantir une disponibilité de 99,7% dans cette situation.

J'espère que ces méthodes aideront le lecteur à trouver des relations quantitatives entre les garanties et leurs conséquences lorsque nous travaillons avec des services dont les sécurités opérationnelles sont interdépendantes.