(12) DEMANDE INTERNATIONALE PUBLIÉE EN VERTU DU TRAITÉ DE COOPÉRATION EN MATIÈRE DE BREVETS (PCT)

(19) Organisation Mondiale de la Propriété Intellectuelle
Bureau international

(43) Date de la publication internationale
04 juillet 2019 (04.07.2019)

WIPO | PCT

(10) Numéro de publication internationale
WO 2019/129986 A1

(51) Classification internationale des brevets :
*G06F 11/20* (2006.01) *G06F 11/14* (2006.01)

(21) Numéro de la demande internationale :
PCT/FR2018/053535

(22) Date de dépôt international :
21 décembre 2018 (21.12.2018)

(25) Langue de dépôt : français

(26) Langue de publication : français

(30) Données relatives à la priorité :
17 63237 27 décembre 2017 (27.12.2017) FR

(71) Déposant : **BULL SAS** [FR/FR] ; rue Jean Jaurès, 78340 LES CLAYES SOUS BOIS (FR).

(72) Inventeurs : **DUGUE, Sébastien** ; Route des Fromenteaux, 38119 SAINT THEOFFREY (FR). **LAFERRIERE, Christophe** ; 1, Chemin des Essarts, 38100 GRENOBLE (FR). **WELTERLEN, Benoit** ; 46, Rue Pierre Sémard, 38000 GRENOBLE (FR).

(74) Mandataire : **CABINET PLASSERAUD** et al. ; 66 rue de la Chaussée d'Antin, 75440 PARIS CEDEX 09 (FR).

(81) États désignés *(sauf indication contraire, pour tout titre de protection nationale disponible)* : AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC,

(54) **Title:** METHOD OF FAULT MANAGEMENT IN A NETWORK OF NODES AND ASSOCIATED PART OF NETWORK OF NODES

(54) **Titre :** PROCEDE DE GESTION DE PANNE DANS UN RESEAU DE NOEUDS ET PARTIE DE RESEAU DE NOEUDS ASSOCIEE

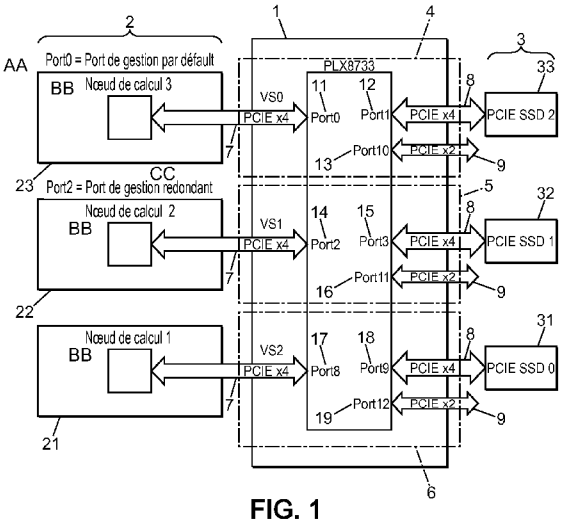


FIG. 1

AA Default management port
BB Calculation node
CC Redundant management port

**(57) Abstract:** The invention relates to a method of fault management in a network of nodes (2), comprising, for each node considered (2) of all or part of the nodes (2) of the network performing one and the same calculation: firstly, a step of local backup of the state of this node considered (21), at the level of a storage medium (31) for this node considered (21), the link (6) between this storage medium (31) and this node considered (21) being able to be redirected from this storage medium (31) to another node (23), thereafter, a step of relaunching: either of the node considered (21) if the latter is not defective, on the basis of the local backup of the state of this node considered (21), or of an operational node (23) different from the node considered (21), if the node considered (21) is defective, on the basis of the recovery of the local backup of the state of this node considered (21), by redirecting said link (6) between the node considered (21) and its storage medium (31) so as to connect said storage medium (31) to said operational node (23), the local backups of these nodes considered (2), used for the relaunching steps, are mutually consistent so as to correspond to one and the same state of this calculation.

**(57) Abrégé :** L'invention concerne un procédé de gestion de panne dans un réseau de nœuds (2), comprenant, pour chaque nœud considéré (2) de tout ou partie des nœuds (2) du réseau effectuant un même calcul : d'abord, une étape de sauvegarde locale de l'état de ce nœud considéré (21), au niveau d'un support de stockage (31) de ce nœud considéré (21), le lien (6) entre ce support de stockage (31) et ce nœud considéré (21) pouvant être redirigé de ce support de stockage (31) vers un autre nœud (23), ensuite, une étape de relance : soit du nœud considéré (21) si celui-ci n'est pas défaillant, à partir de la sauvegarde locale de l'état de ce nœud considéré

WO 2019/129986 A1

*[Suite sur la page suivante]*

SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **États désignés** *(sauf indication contraire, pour tout titre de protection régionale disponible)* : ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), eurasien (AM, AZ, BY, KG, KZ, RU, TJ, TM), européen (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

**Publiée:**

— *avec rapport de recherche internationale (Art. 21(3))*

(21), soit d'un nœud opérationnel (23) différent du nœud considéré (21), si le nœud considéré (21) est défaillant, à partir de la récupération de la sauvegarde locale de l'état de ce nœud considéré (21), en redirigeant ledit lien (6) entre le nœud considéré (21) et son support de stockage (31) de manière à relier ledit support de stockage (31) vers ledit nœud opérationnel (23), les sauvegardes locales de ces nœuds considérés (2), utilisées pour les étapes de relance, sont cohérentes entre elles de sorte à correspondre à un même état de ce calcul.

# PROCEDE DE GESTION DE PANNE DANS UN RESEAU DE NŒUDS ET PARTIE DE RESEAU DE NŒUDS ASSOCIEE

## DOMAINE DE L'INVENTION

L'invention concerne le domaine des procédés de gestion de panne dans un réseau de nœuds et le domaine des parties de réseau de nœuds associées à cette gestion de panne.

## CONTEXTE DE L'INVENTION

Dans un réseau de nœuds effectuant un même calcul, des sauvegardes sont effectuées à un ou à plusieurs niveaux, ce sont des sauvegardes multiniveaux. Lorsqu'une panne survient, le calcul peut être récupéré au moins en partie sans devoir être intégralement recommencé, justement grâce aux sauvegardes effectuées. Selon le type de panne survenu, l'un ou l'autre des niveaux de sauvegarde est utilisé pour récupérer le calcul en partie ou en même en majorité ou encore presque totalement.

Les applications distribuées peuvent durer beaucoup plus longtemps que la durée moyenne sans panne d'un réseau, encore appelée MTBF d'un cluster (pour « Mean Time Between Failures » en langue anglaise), elles ont donc beaucoup de chances d'être interrompues. Elles n'ont en général pas de solution interne de gestion des pannes, ce peut alors entraîner d'abord la perte des données de sauvegarde locales en cas de panne physique du nœud de calcul, ensuite suivie de la perte de l'ensemble de l'avancement du calcul entraînée par la perte d'un seul nœud de calcul mais dont les données locales de sauvegarde ne peuvent plus être récupérées.

Il existe des solutions de sauvegarde et de redémarrage (« checkpoint/restart » en langue anglaise) à partir de sauvegarde permettant

aux applications de sauvegarder leur contexte régulièrement sur différents niveaux de sauvegarde, plus ou moins rapides.

Les différents niveaux de sauvegarde vont de la sauvegarde la plus locale à la sauvegarde la plus globale, allant ainsi de la sauvegarde la plus simple et la plus rapide à la sauvegarde la plus complexe, la plus lente et la plus coûteuse, allant également ainsi de la sauvegarde la plus fragile et la plus faible à la sauvegarde la plus robuste et la plus résiliente.

Selon un art antérieur correspondant à la librairie FTI (« interface de tolérance aux pannes » ou bien « fault tolerance interface » en langue anglaise), il est connu quatre niveaux de sauvegarde qui sont :

- Le premier niveau L1 qui réalise une sauvegarde locale, simple et peu coûteuse, effectuée très souvent, d'où une perte minime en temps de calcul lors d'une panne pouvant être récupérée à ce seul premier niveau L1,
- Le deuxième niveau L2 qui réalise une première sauvegarde intermédiaire par duplication sur un nœud partenaire, moins simple et plus coûteuse, d'où une perte plus importante en temps de calcul lors d'une panne ne pouvant être récupérée qu'à ce deuxième niveau L2,
- Le troisième niveau L3, qui réalise une deuxième sauvegarde intermédiaire par encodage de type Reed-Solomon, encore moins simple et encore plus coûteuse, d'où une perte encore un peu plus importante en temps de calcul lors d'une panne ne pouvant être récupérée qu'à ce troisième niveau L3,
- Le quatrième niveau L4, qui réalise une sauvegarde globale à un niveau système de fichier, complexe et très coûteuse, d'où une perte vraiment importante en temps de calcul lors d'une panne ne pouvant être récupérée qu'à ce quatrième niveau L4.

Du niveau local L1 au niveau global L4, la sauvegarde est de plus en plus robuste et résiliente, mais elle devient également de plus en plus complexe et coûteuse. C'est pour cela, que la sauvegarde du premier niveau

L1 est réalisée souvent, la sauvegarde du deuxième niveau L2 moins souvent, la sauvegarde du troisième niveau L3 encore moins souvent, la sauvegarde du quatrième niveau L4 relativement rarement. Par conséquent, statistiquement, lorsqu'une panne survient, le dernier état cohérent qui peut être récupéré est, très récent au premier niveau L1, moins récent au deuxième niveau L2, encore moins récent au troisième niveau L3, plus ancien au quatrième niveau L4. Par conséquent le volume de travail perdu est, très faible au premier niveau L1, relativement limité au deuxième niveau L2, notable au troisième niveau L3, plus important ancien au quatrième niveau L4.

## RESUME DE L'INVENTION

Le but de la présente invention est de fournir un procédé de gestion de panne dans un réseau de nœuds palliant au moins partiellement les inconvénients précités.

Plus particulièrement, l'invention vise à fournir un procédé de gestion de panne dans un réseau de nœuds améliorant le compromis entre efficacité d'une part et coût et complexité d'autre part, pour au moins un niveau de sauvegarde considéré.

Plus particulièrement, l'invention vise à fournir un procédé de gestion de panne dans un réseau de nœuds qui ait une efficacité similaire ou comparable à celle d'une sauvegarde de niveau intermédiaire, préférentiellement celle de la première sauvegarde intermédiaire, avantageusement celle du deuxième niveau L2, pour un coût et une complexité similaires ou comparables à ceux d'une sauvegarde de niveau local, avantageusement ceux du premier niveau L1.

Pour cela, l'invention propose de rendre le lien entre un support de stockage et son nœud redirigeable vers un autre nœud, de manière à pouvoir disposer de la sauvegarde réalisée sur le support de stockage lorsque le nœud est défaillant, éventuellement par le biais d'une copie au niveau d'un

nœud voisin, mais sans avoir réalisé de copie de sauvegarde au niveau d'un nœud voisin pour la plupart des nœuds qui ne sont pas défaillants ou pour tous les nœuds qui ne sont pas défaillants. Le lien entre un support de stockage et son nœud n'est pas réalisé directement, mais indirectement au travers d'un élément de réseau apte à reconfigurer ce lien pour rattacher ce support de stockage à un autre nœud lorsque le précédent nœud devient défaillant. Cet élément de réseau relie plusieurs nœuds à leurs supports de stockage respectifs, chaque nœud étant relié à son support de stockage associé (ou éventuellement à ses supports de stockage associés).

En résumé, pour un nœud défaillant, on dispose d'une sauvegarde malgré la défaillance de ce nœud, d'où une efficacité similaire au deuxième niveau L2, mais on ne réalise pas d'opération plus complexe qu'une simple sauvegarde locale pour la plupart ou pour tous les nœuds qui ne sont pas défaillants, ce qui est la majorité ou même la grande majorité des nœuds du réseau effectuant le calcul considéré, d'où un coût et une complexité du moins comparable sinon similaire à celles du premier niveau L1.

Ainsi, selon des modes de réalisation de l'invention, le coût du deuxième niveau L2 est économisé, alors qu'est conservée la capacité à redémarrer une application à partir des sauvegardes du premier niveau L1 en cas de panne d'un nœud de calcul. La copie vers un nœud voisin traditionnellement effectuée au deuxième niveau L2 n'est ici pas réalisée pendant l'exécution de l'application de manière préventive pour tous les nœuds de calcul, mais seulement en cas de panne et seulement pour les nœuds défaillants après survenue de la panne. La copie est alors faite uniquement pour relancer l'application avec les données manquantes, remontées de la sauvegarde locale.

D'une part, cette sauvegarde de coût et de complexité similaires ou comparables à une sauvegarde locale présente l'efficacité d'une sauvegarde intermédiaire, ce qui améliore beaucoup le rapport qualité prix de cette sauvegarde.

D'autre part, cette sauvegarde de coût et de complexité similaires ou comparables à une sauvegarde locale présentant l'efficacité d'une sauvegarde intermédiaire, permet préférentiellement de remplacer à la fois une sauvegarde locale classique et une ou plusieurs sauvegardes intermédiaires classiques, en ne gardant en plus que la sauvegarde globale en dernier recours, pour gérer les pannes les plus sévères, la plupart des pannes pouvant être maintenant gérées par la sauvegarde locale-intermédiaire proposée par l'invention, celle-ci ayant l'efficacité d'une sauvegarde intermédiaire pratiquement au prix d'une sauvegarde locale.

A cette fin, la présente invention propose un procédé de gestion de panne dans un réseau de nœuds, comprenant, pour chaque nœud considéré de tout ou partie des nœuds du réseau effectuant un même calcul : d'abord, une étape de sauvegarde locale de l'état de ce nœud considéré, au niveau d'un support de stockage de ce nœud considéré, le lien entre ce support de stockage et ce nœud considéré pouvant être redirigé de ce support de stockage vers un autre nœud, ensuite, une étape de relance : soit du nœud considéré si celui-ci n'est pas défaillant, à partir de la sauvegarde locale de l'état de ce nœud considéré, soit d'un nœud opérationnel différent du nœud considéré, si le nœud considéré est défaillant, à partir de la récupération de la sauvegarde locale de l'état de ce nœud considéré, en redirigeant ledit lien entre le nœud considéré et son support de stockage de manière à relier ledit support de stockage vers ledit nœud opérationnel, les sauvegardes locales de ces nœuds considérés, utilisées pour les étapes de relance, sont cohérentes entre elles de sorte à correspondre à un même état de ce calcul.

A cette fin, la présente invention propose également une partie d'un réseau de nœuds, comprenant : un commutateur, une pluralité de nœuds d'un groupe de nœuds effectuant un même calcul dans ce réseau de nœuds, plusieurs supports de stockage respectivement rattachés à ces nœuds par l'intermédiaire du commutateur, ces supports de stockage étant structurés et disposés pour sauvegarder localement l'état des nœuds auxquels ils sont respectivement rattachés, ce commutateur étant structuré et disposé pour

changer au moins un rattachement entre un nœud défaillant et son support de stockage lequel va alors devenir rattaché à un autre nœud de ladite pluralité de nœuds.

Selon des modes de réalisation de l'invention, ce procédé de gestion de panne permet l'optimisation et la résilience de la sauvegarde de données d'une application susceptible de rencontrer une panne et le redémarrage de cette application à l'aide des données sauvegardées.

Selon des modes de réalisation de l'invention, ce procédé de gestion de panne offre la possibilité de migrer un support de stockage du nœud de calcul en panne grâce au commutateur, permettant alors d'exploiter directement les données locales sauvegardées d'un nœud de calcul en panne en un temps qui ne dépend que peu de la quantité de données sauvegardées.

Selon des modes de réalisation de l'invention, ce procédé de gestion de panne offre une tolérance aux pannes qui s'étend dans l'absolu à un grand nombre de nœuds de calcul tant qu'il existe, sur chaque nœud de calcul qui tombe en panne, un nœud de calcul rattaché au même commutateur qui pourra reprendre le calcul du nœud de calcul tombé en panne. Il est alors possible de repartir de la dernière sauvegarde locale, même si un grand nombre des nœuds de calcul tombe en panne, tant que chaque nœud de calcul tombé en panne dispose d'un commutateur et d'un nœud de calcul voisin non défaillant connecté au même commutateur.

Suivant des modes de réalisation préférés, l'invention comprend une ou plusieurs des caractéristiques suivantes qui peuvent être utilisées séparément ou en combinaison partielle entre elles ou en combinaison totale entre elles, appliquées à l'un ou à l'autre des objets de l'invention précités.

De préférence, pour chaque nœud considéré non défaillant : il n'y a pas d'étape de récupération par un autre nœud, de la sauvegarde locale de l'état de ce nœud considéré non défaillant.

Ainsi, pour tous les nœuds non défaillants, c'est-à-dire pour la grande majorité des nœuds du réseau effectuant un même calcul, aucune opération plus complexe qu'une simple sauvegarde locale n'a été réalisée.

De préférence, le nœud opérationnel différent du nœud considéré défaillant est un nœud de rechange.

Ainsi, les nœuds non défaillants, qui ont déjà chacun leur propre tâche de calcul à effectuer, ne sont pas chargés en plus d'effectuer cette opération de récupération d'un nœud voisin défaillant. Une éventuelle surcharge de ces nœuds non défaillants est alors évitée.

De préférence, ladite redirection dudit lien entre le nœud considéré et son support de stockage de manière à relier ledit support de stockage vers ledit nœud opérationnel, est réalisée par un changement d'aiguillage dans un commutateur reliant plusieurs nœuds à leurs supports de stockage.

Ainsi, cette redirection est effectuée par une opération simple réalisée au niveau d'un élément de réseau qui est fiable et maîtrisé.

De préférence, toutes les étapes de relance de nœuds sont synchronisées entre elles, de manière à relancer tous lesdits nœuds dans un même état de calcul.

Ainsi, une complète cohérence du reste du calcul, effectué après la relance, est assurée.

De préférence, dans une première réalisation, l'étape de récupération comprenant une sous-étape de migration, vers le nœud opérationnel ou vers le nœud de rechange, du support de stockage de la sauvegarde locale de l'état du nœud défaillant, qui rattache, ce support de stockage de la sauvegarde locale de l'état du nœud défaillant, à un endroit prédéterminé de l'arborescence de fichiers de ce nœud opérationnel ou de ce nœud de rechange, ce nœud opérationnel ou ce nœud de rechange venant ensuite lire, à cet endroit prédéterminé, le lien vers la sauvegarde locale de l'état du nœud défaillant, lors de l'étape de relance de ce nœud opérationnel ou de ce nœud de rechange, aucune copie, de la sauvegarde locale de l'état du nœud défaillant, n'étant réalisée, au niveau d'un autre nœud.

Ainsi, la récupération est rendue encore plus simple, puisque même pour un nœud défaillant, il n'y a aucune duplication de sa sauvegarde locale

qui est réalisée. Il n'y avait déjà pas de duplication réalisée pour la sauvegarde locale des nœuds non défaillants.

De préférence, le nœud de rechange et le nœud défaillant qu'il remplace appartiennent tous les deux à la même lame de calcul.

Ainsi, la redirection du lien entre support et nœud est effectuée plus simplement et plus rapidement, en raison de la proximité géographique entre eux du nœud de rechange et du nœud défaillant.

De préférence, dans une deuxième réalisation, l'étape de récupération comprend : d'abord, une sous-étape de migration, vers un nœud intermédiaire, du support de stockage de la sauvegarde locale de l'état du nœud défaillant, qui rattache, ce support de stockage de la sauvegarde locale de l'état du nœud défaillant, à un endroit prédéterminé de l'arborescence de fichiers de ce nœud intermédiaire, ensuite, une sous-étape de copie, de la sauvegarde locale de l'état du nœud défaillant, du support de stockage rattaché au nœud intermédiaire vers le support de stockage du nœud opérationnel ou du nœud de rechange, ce nœud opérationnel ou ce nœud de rechange venant ensuite lire son support de stockage, lors de son étape de relance.

Ainsi, même si la récupération, pour un nœud défaillant, réalise une duplication de sa sauvegarde locale, en revanche il n'y a pas besoin d'indiquer au nœud opérationnel ou au nœud de rechange à quel endroit aller chercher le lien vers le support de stockage du nœud défaillant à récupérer. Par contre, bien sûr, il n'y a toujours pas de duplication réalisée pour la sauvegarde locale des nœuds non défaillants.

La sous-étape de copie n'est pas nécessaire et peut être évitée en remontant le support de stockage au bon endroit dans l'arborescence de fichier du nœud de rechange comme dans la première réalisation décrite précédemment, ou bien en changeant simplement la configuration de la bibliothèque de tolérance aux pannes. Cette remontée est instantanée et permet alors d'éviter une copie supplémentaire qui pourrait prendre un peu de temps en fonction du volume de données à sauvegarder.

De préférence, pour tout ou partie des nœuds du réseau effectuant un même calcul : le nœud de rechange et le nœud défaillant qu'il remplace appartiennent à des lames de calcul différentes.

Ainsi, la redirection du lien entre support et nœud peut être effectuée même en cas de panne importante et même relativement généralisée au niveau de toute une lame de calcul.

De préférence, toutes ces étapes sont réalisées pour tous les nœuds du réseau effectuant un même calcul.

Ainsi, le bénéfice du procédé de gestion de panne proposé par l'invention, est généralisé en étant étendu à l'ensemble des nœuds du réseau effectuant un même calcul.

De préférence, la sous-étape de migration change le rattachement du support de stockage de la sauvegarde locale de l'état du nœud défaillant en passant par l'intermédiaire d'un commutateur auquel était rattaché le nœud défaillant et son support de stockage de la sauvegarde locale de l'état du nœud défaillant, mais sans passer par le nœud défaillant lui-même.

Ainsi, la redirection peut être effectuée même en cas de panne physique complète du nœud défaillant.

De préférence, le changement de rattachement est réalisé par l'envoi d'une commande au commutateur, cette commande passant par l'un des nœuds rattachés au commutateur par un port de management.

Ainsi, c'est le port de management ou de gestion qui est affecté à la récupération des supports de stockage des nœuds défaillants rattachés à un même commutateur.

De préférence, ce commutateur est un commutateur PCIe (« Peripheral Component Interconnect express » en langue anglaise).

Ainsi, ce commutateur est particulièrement avantageux car il est particulièrement adapté à pouvoir faire communiquer entre eux des périphériques sans devoir passer par l'intermédiaire d'un microprocesseur, donc en pouvant contourner un nœud de calcul défaillant par exemple.

L'utilisation de ce commutateur PCIe permet de rattacher le support de stockage, par exemple un disque de stockage, contenant les sauvegardes locales du nœud de calcul en panne à un nœud de calcul de rechange. Cette opération est rapide et ne nécessite pas de copie systématique des données de sauvegarde locales, et en particulier pas pour les nœuds de calcul non défaillants.

De préférence, 3 à 10 nœuds sont rattachés à un même commutateur.

Ainsi, le commutateur peut gérer facilement ce petit groupe de nœuds pour lequel un seul nœud de rechange semble devoir être suffisant.

De préférence, le procédé de gestion de panne comprend aussi, pour tout ou partie des nœuds du réseau effectuant un même calcul : une étape de sauvegarde globale de l'ensemble de ces nœuds, réalisée moins souvent que l'ensemble des étapes de sauvegarde locales de ces nœuds.

Ainsi, avec d'une part la sauvegarde locale-intermédiaire proposée par l'invention pour de gérer la grande majorité des pannes de manière simple et efficace, et d'autre part la sauvegarde globale plus complexe et plus coûteuse mais réservée à une minorité de pannes sévères, un excellent compromis entre complexité globale et efficacité globale du procédé de gestion de panne proposé par l'invention est réalisé.

Dans le cadre d'une application tolérante aux pannes, en utilisant plusieurs niveaux de sauvegarde, comme ici une sauvegarde locale rapide et une sauvegarde globale distante plus complexe et plus coûteuse, le procédé de gestion de panne proposé par l'invention permet alors de redémarrer l'application, suite à une panne physique sur un nœud, même complète, et ceci dans la plupart des cas, en repartant de toutes les sauvegardes locales, plus récentes et moins coûteuses, au lieu de devoir repartir des sauvegardes distantes qui sont très souvent nettement plus anciennes, quelques cas plus rares de panne nécessitant parfois l'utilisation de la sauvegarde globale distante. La possibilité de récupérer les données locales du nœud de calcul en panne permet de redémarrer dans la plupart des cas l'application à partir des sauvegardes locales les plus récentes.

De préférence, pour tout ou partie des nœuds du réseau effectuant un même calcul : le réseau comprend entre 1 et 5 nœuds de rechange pour 100 nœuds effectuant ledit même calcul.

Ainsi, le surdimensionnement de la taille du réseau est minime, tandis que la grande majorité des pannes ordinaires sera gérée de manière très efficace.

De préférence, pour tout ou partie des nœuds du réseau effectuant un même calcul : ce procédé de gestion de panne ne comprend aucun autre type de sauvegarde locale de l'état de ces nœuds.

Ainsi, la simplicité et le faible coût de ce type de sauvegarde locale-intermédiaire proposée par l'invention est d'une simplicité similaire à celle d'une sauvegarde purement locale, pour une efficacité bien supérieure.

De préférence, pour tout ou partie des nœuds du réseau effectuant un même calcul : les supports de stockage sont des mémoires flash.

Ainsi, les mémoires utilisées sont simples, rapides et permanentes.

De préférence, ces mémoires flash sont des mémoires NVMe (« Non Volatile Memory express » en langue anglaise).

Ainsi, les mémoires utilisées sont particulièrement bien adaptées pour communiquer avec un commutateur PCIe (« Peripheral Component Interconnect express » en langue anglaise).

De préférence, ce commutateur étant structuré et disposé pour changer au moins un rattachement entre un nœud défaillant et son support de stockage, sur commande extérieure au commutateur.

Ainsi, c'est le commutateur qui effectue la redirection du lien entre le support de stockage et son nœud défaillant, dès qu'un élément externe au commutateur a mis en évidence cette défaillance de nœud et l'a prévenu.

De préférence, ladite pluralité de nœuds comprend entre 3 et 10 nœuds.

Ainsi, le commutateur peut gérer facilement ce petit groupe de nœuds pour lequel un seul nœud de rechange semble devoir être suffisant.

De préférence, ce commutateur est un commutateur PCIe.

Ainsi, ce commutateur est particulièrement avantageux car il est particulièrement adapté à pouvoir faire communiquer entre eux des périphériques sans devoir passer par l'intermédiaire d'un microprocesseur.

De préférence, les supports de stockage sont des mémoires flash, de préférence des mémoires NVMe.

Ainsi, les mémoires utilisées sont particulièrement bien adaptées pour communiquer avec un commutateur PCIe.

Préférentiellement, le réseau de nœuds de calcul comprend au moins 1000 nœuds de calcul, avantageusement au moins 5000 nœuds de calcul, encore plus avantageusement au moins 10000 nœuds de calcul, rendant d'autant plus intéressant le procédé de gestion de panne selon l'invention car la perte complète d'un calcul en cours devient alors d'autant plus critique que le réseau est grand.

Le principe de tolérance aux pannes est d'autant plus important qu'une application s'exécute sur un cluster qui est constitué d'un plus grand nombre de nœuds de calcul. Plus le nombre de processeurs, mémoires et autres périphériques, est important, plus la probabilité qu'une panne survienne avant la fin de l'exécution est grande. Les applications qui ont pour but de s'exécuter sur ce type de plateforme vont utiliser des librairies de tolérance aux pannes qui leur permettent de sauvegarder (« checkpoint » en langue anglaise) les données nécessaires à un redémarrage (« restart » en langue anglaise) dans un état le plus proche possible de l'état qui existait juste avant la panne. Ainsi, ces applications ne sont pas contraintes de recommencer le calcul depuis le début.

D'autres caractéristiques et avantages de l'invention apparaîtront à la lecture de la description qui suit d'un mode de réalisation préféré de l'invention, donnée à titre d'exemple et en référence aux dessins annexés.

## BREVE DESCRIPTION DES DESSINS

La figure 1 représente schématiquement un exemple, d'une partie de réseau incluant un groupe de nœuds et leurs supports de stockage reliés entre eux par un commutateur PCIe, selon un mode de réalisation de l'invention, au niveau de laquelle partie de réseau, un exemple de procédé de gestion de panne selon l'invention peut se dérouler.

## DESCRIPTION DETAILLEE DE L'INVENTION

La figure 1 représente schématiquement un exemple, d'une partie de réseau incluant un groupe de nœuds et leurs supports de stockage reliés entre eux par un commutateur PCIe, selon un mode de réalisation de l'invention, au niveau de laquelle partie de réseau, un exemple de procédé de gestion de panne selon l'invention peut se dérouler.

Cette partie de réseau comprend plusieurs nœuds de calcul 2, trois nœuds de calcul 21, 22 et 23, dans l'exemple de la figure 1, ainsi que plusieurs supports de stockage 3, trois supports de stockage 31, 32 et 33, dans l'exemple de la figure 1.

Ces nœuds de calcul 2 et leurs supports de stockage 3 forment un groupe de nœuds de calcul géré par un commutateur PCIe 1 reliant ces nœuds de calcul 2 à leurs supports de stockage 3 respectifs, par l'intermédiaire de connexions bidirectionnelles PCIe 7, 8 ou 9. Ces connexions PCIe 7, 8 ou 9, peuvent être des connexions PCIe à plusieurs voies. La connexion 7 est une connexion à 4 voies. La connexion 8 est une connexion à 4 voies. La connexion 9 est une connexion à 2 voies, ici non encore utilisée, chaque connexion 9 étant d'un côté rattachée à l'un des ports 13, 16 ou 19 et restant libre de l'autre côté. Chaque connexion PCIe 7 relie respectivement l'un des nœuds de calcul 21 à 23 à l'un des ports 0, 2 ou 8 (numérotation du commutateur électronique PLX 8733, mais un autre

Switch PCIe peut être utilisé), respectivement référencés 11, 14 ou 17 sur la figure 1, du commutateur 1. Chaque connexion PCIe 8 relie respectivement l'un des supports de stockage 31 à 33 à l'un des ports 1, 3 ou 9, respectivement référencés 12, 15 ou 18 sur la figure 1, du commutateur 1. La connexion 7, les ports 0 et 1, respectivement référencés 11 et 12 sur la figure 1, du commutateur 1, la connexion 8 forment ensemble un lien 4 reliant le nœud de calcul 23 à son support 33. La connexion 7, les ports 2 et 3, respectivement référencés 14 et 15 sur la figure 1, du commutateur 1, la connexion 8 forment ensemble un lien 5 reliant le nœud de calcul 22 à son support 32. La connexion 7, les ports 8 et 9, respectivement référencés 17 et 18 sur la figure 1, du commutateur 1, la connexion 8 forment ensemble un lien 6 reliant le nœud de calcul 21 à son support 31. Les connexions PCIe 7, 8 ou 9, peuvent être regroupées sur un bus PCIe.

Le nœud de calcul 23 est rattaché au port de gestion ou de management par défaut, c'est-à-dire que c'est par lui que transite les envois vers l'extérieur du groupe de nœuds de calcul 2 et les réceptions en provenance de l'extérieur du groupe de nœuds de calcul 2. En cas de défaillance de ce nœud de calcul 23, celui-ci est remplacé par le nœud de calcul 22 qui est rattaché au port de gestion ou de management redondant, lequel nœud de calcul 22 devient alors rattaché au nouveau port de gestion ou de management effectif.

Lorsqu'un nœud de calcul tombe en panne physique, par exemple le nœud de calcul 21 ou le nœud de calcul 22, considérons ici le nœud de calcul 21, la dernière sauvegarde locale récente de son état de calcul est stockée sur son support de stockage 31.

Dans un système selon l'art antérieur, le support de stockage 31 n'étant accessible que par son nœud de calcul 21 et celui-ci étant en panne physique complète, cette sauvegarde locale récente deviendrait inaccessible, et il faudrait alors recourir à d'autres niveaux de sauvegarde plus complexes et moins récents, d'où une perte importante d'efficacité globale pour le réseau informatique.

Dans le système selon un mode de réalisation de l'invention, présenté à la figure 1, le rattachement du support de stockage 31 est reconfiguré, c'est-à-dire que le support de stockage 31 va cesser d'être relié à son nœud de calcul 21 de rattachement par défaut, mais va devenir relié au nœud de calcul 23, lequel étant rattaché au port de gestion par défaut, va pouvoir faire remonter la sauvegarde locale de l'état de calcul du nœud de calcul 21 défaillant, depuis le support de stockage 31, vers un autre nœud de calcul de rechange extérieur au groupe de nœuds de calcul 21 à 23, cet autre nœud de calcul reprenant alors à son compte la tâche de calcul interrompue au niveau du nœud de calcul 21 défaillant, à partir de la sauvegarde locale de l'état de calcul du nœud de calcul 21 remontée depuis le support de stockage 31.

Dans le commutateur 1, le port 9 (numérotation du commutateur électronique PLX8733, mais un autre Switch PCIe peut être utilisé) référencé 18 (sur la figure 1), au lieu de rester relié en permanence au port 8 référencé 17 comme avant la défaillance du nœud de calcul 21, va être, au moins temporairement relié au port 0 référencé 11, afin de permettre au nœud de calcul 23 de venir lire, dans le support de stockage 31, les données sauvegardées représentatives de l'état de calcul du nœud de calcul 21, juste avant ou peu de temps avant, sa défaillance. Ainsi, la sauvegarde locale, dans le support de stockage 31, de l'état de calcul du nœud de calcul 21 avant sa défaillance, va pouvoir remonter jusqu'au nœud de calcul 23 et être ainsi exploitée pour relancer le calcul avec un très bon compromis entre simplicité de la sauvegarde et efficacité de la relance.

Alternativement, si le nœud de calcul 23 est lui-même un nœud de rechange, il peut reprendre à son compte la tâche de calcul interrompue par le nœud de calcul 21 défaillant.

Si c'est le nœud 23 qui devient défaillant, il est d'abord remplacé par le nœud de calcul 22 comme rattaché au port de management, et le nœud de calcul 22, comme rattaché au nouveau port de management, réalise les opérations auparavant réalisés par le nœud de calcul 23 si celui-ci n'était pas devenu défaillant.

La gestion du commutateur 1 est maintenant décrite par un scenario manuel clarifiant et expliquant les différentes opérations à effectuer, comme par exemple la migration du support de stockage 31 d'un nœud de calcul défaillant 21 à un nœud de calcul de rechange 23, le transfert des données, le redémarrage de l'application. La gestion des données côté applicatif est abstraite par la librairie FTI.

L'application exécutée est fournie dans les exemples de la librairie FTI : hdf.exe. Cette application est lancée sur deux nœuds de calcul. Elle va réaliser des sauvegardes locales, sur les disques de stockage NVMe reliés à ces nœuds de calcul par le commutateur PCIe à intervalles réguliers ainsi qu'une sauvegarde globale sur un serveur NFS (« Network File System » en langue anglaise) de façon moins fréquente. Une fois l'application lancée, une panne est générée sur un des deux nœuds de calcul. La migration du disque de stockage NVMe du nœud défaillant vers un autre nœud va alors permettre le transfert des données de la dernière sauvegarde locale de ce nœud défaillant vers le nœud de rechange. Une fois ce transfert réalisé, l'application peut être relancée et reprend le calcul à la dernière sauvegarde locale des deux nœuds de calcul au lieu de la dernière sauvegarde globale plus ancienne de la partie de réseau.

Dans ce contexte, une reprise sur panne est réalisée avec les données locales de sauvegarde d'une application MPI (« Message Passing Interface » en langue anglaise) d'un nœud de calcul tombant en panne. Les données sont récupérées grâce à la migration du support de stockage du nœud de calcul défaillant sur un nœud de calcul voisin de la même lame de calcul. Ces données sont ensuite transmises à un second nœud de rechange qui reprendra le calcul. L'intérêt de cette reprise sur panne est qu'elle permet à l'application de redémarrer à partir des sauvegardes locales de tous les nœuds de calcul. Ces sauvegardes locales moins coûteuses sont également plus récentes la plupart du temps, et ne sont au pire qu'aussi récentes, que les sauvegardes globales. De plus, la quantité de données transférées afin de

reprendre le calcul va être bien plus faible que dans le cas d'un redémarrage à partir d'une sauvegarde globale.

Dans le développement qui suit, les parties de texte dans un encadré ou bien entre crochets concernent des lignes de code informatique.

La configuration de la librairie FTI a été modifiée [« config.fti » dans le répertoire « examples »], afin d'autoriser l'exécution sur deux nœuds de calcul différents :

```
[basic]
head                = 0
node_size           = 1
ckpt_dir            = /localckpt/
glbl_dir            = /globalckpt/
meta_dir            = /home/user/.fti
ckpt_l1             = 1
ckpt_l2             = 2
ckpt_l3             = 3
ckpt_l4             = 4
inline_l2           = 1
inline_l3           = 1
inline_l4           = 1
keep_last_ckpt      = 0
group_size          = 2
verbosity           = 1

[restart]
failure             = 1
exec_id             = 2016-10-27_15-25-56
[injection]
rank                = 0
number              = 0
position            = 0
frequency           = 0[advanced]

block_size          = 1024
mpi_tag             = 2612
local_test          = 1
```

Sur chacun des deux nœuds de calcul, appelés ici N1 (référencé 21 sur la figure 1) et N2 (référencé 22 sur la figure 1), le répertoire de sauvegarde locale [« /localckpt/ »] est un montage du disque de stockage SSD (« Solid State Disk » en langue anglaise) disponible sur chaque nœud de calcul, le disque de stockage S0 (référencé 31 sur la figure 1) sur le nœud de calcul N1 et le disque de stockage S1 (référencé 32 sur la figure 1) sur le nœud de calcul N2 :

```
$ mpirun -host N1,N2 --mca pml ob1 --mca btl self,vader,tcp --mca coll ^hcoll,fca --mca oob_tcp_if_include enp12s0f0 --mca btl_tcp_if_include enp12s0f0 -n 2 hdf.exe
[ FTI  Information ] : Reading FTI configuration file (config.fti)...
[ FTI  Information ] : The execution ID is: 2016-11-09_17-09-10
FTI_Topo->nodeSize : 1 , FTI_Topo->nbProc : 2
FTI_Topo->nbNodes : 2 FTI_Topo->nbNodes : 2 FTI_Topo->groupSize : 2 , FTI_Topo->nodeSize 1
FTI_Topo->nodeSize : 1 , FTI_Topo->nbProc : 2
FTI_Topo->nbNodes : 2 FTI_Topo->nbNodes : 2 FTI_Topo->groupSize : 2 , FTI_Topo->nodeSize 1
[ FTI  Information ] : FTI has been initialized.
Local data size is  2048 x  1027 =   32. MB (   32).
Target precision :   0.00500
[ FTI  Information ] : Variable ID 0 to protect. Current ckpt. size per rank is 0.00MB.
[ FTI  Information ] : Variable ID 2 to protect. Current ckpt. size per rank is 16.05MB.
[ FTI  Information ] : Variable ID 1 to protect. Current ckpt. size per rank is 32.09MB.
Step :   100, error =   0.25488
[ FTI  Information ] : Post-checkpoint took 0.00 sec. (Ag:0.00s, Pt:0.00s, Cl:0.00s)
[ FTI  Information ] : Ckpt. ID 1 (L1) (32.09 MB/proc) taken in 0.13 sec. (Wt:0.00s, Wr:0.12s, Ps:0.00s)
Step :   200, error =   0.12412
[FTI Error - 000000] : Cannot rename local directory : No such file or directory
[FTI Error - 000001] : Cannot rename local directory : No such file or directory
[ FTI  Information ] : Post-checkpoint took 0.12 sec. (Ag:0.00s, Pt:0.00s, Cl:0.12s)
[ FTI  Information ] : Ckpt. ID 2 (L4) (32.09 MB/proc) taken in 0.48 sec. (Wt:0.00s, Wr:0.36s, Ps:0.12s)
Step :   300, error =   0.08203
Step :   400, error =   0.06127
[ FTI  Information ] : Post-checkpoint took 0.00 sec. (Ag:0.00s, Pt:0.00s, Cl:0.00s)
[ FTI  Information ] : Ckpt. ID 3 (L1) (32.09 MB/proc) taken in 0.13 sec. (Wt:0.00s, Wr:0.12s, Ps:0.00s)
```

Une panne du nœud de calcul N1 est provoquée, causant alors une interruption du calcul de ce nœud de calcul N1 :

[$ ipmitool –H bmc-N1 –Uuser -Ppass power off]

Dans la trace précédente, les sauvegardes locales de premier niveau L1 sont différenciées des sauvegardes globales de quatrième niveau L4. Le calcul a été interrompu après avoir fait une sauvegarde globale de quatrième niveau L4 et une sauvegarde locale de premier niveau L1 plus récente que la sauvegarde globale de quatrième niveau L4, ce qui statistiquement correspondra à la grande majorité, pour ne pas dire à la quasi-totalité, des cas de survenue de panne en pratique.

Le nœud de calcul N1 étant considéré en panne, le disque de stockage S0 qui lui était attaché va être migré vers un autre nœud de calcul, ici le nœud de calcul N3 (référencé 23 sur la figure 1). Le nœud de calcul N3 est le nœud qui possède le port d'administration du commutateur.

La première étape est d'envoyer la commande qui permet de router à nouveau la remise à zéro PCI (« reset » en langue anglaise), le dernier argument étant le numéro du nœud fautif ou défaillant :

[$ ipmitool -Hpm-bmc-N3 -Usuper -Ppass raw 0x3a 0xcd 1]

Cette commande peut être exécutée des nœuds voisins ou du nœud de management ou de gestion.

L'état des ports du commutateur (référencé 1 sur la figure 1) est vérifié :

```
$ bpsm_info
Active management port found @ 01:00:00

Management port control:
  Active mngt port:      0 (enabled)
  Redundant mngt port:   2 (enabled)

Virtual switch enable:   00000007
VS0:
      Upstream port:    0
      Downstream ports: 1 10

VS1:
      Upstream port:    2
      Downstream ports: 3 11

VS2:
      Upstream port:    8
      Downstream ports: 9 12
```

Le port 9 (référencé 18 sur la figure 1) du commutateur, sur lequel se trouve le disque de stockage S0 du nœud de calcul fautif ou défaillant N1 au nœud de calcul courant N3 :

```
$ bpsm_grab_port 9
Sucessfully grabbed downstream port 9

$ bpsm_info
Active management port found @ 01:00:00

Management port control:
  Active mngt port:      0 (enabled)
  Redundant mngt port:   2 (enabled)

Virtual switch enable:   00000007
VS0:
      Upstream port:    0
      Downstream ports: 1 9 10

VS1:
      Upstream port:    2
      Downstream ports: 3 11

VS2:
      Upstream port:    8
      Downstream ports: 12
```

Il est demandé au système de scanner à nouveau le bus PCIe :

```
$ echo 1 > /sys/bus/pci/devices/0000:00:00.0/rescan
```

Le disque de stockage S0 est désormais vu :

```
$ lsblk
NAME        MAJ:MIN RM   SIZE RO TYPE MOUNTPOINT
loop0         7:0    0   1,1G  1 loop
loop1         7:1    0    5G  1 loop
├─live-rw   253:0    0    5G  0 dm   /
└─live-base 253:1    0    5G  1 dm
loop2         7:2    0   512M  0 loop
└─live-rw   253:0    0    5G  0 dm   /
nvme0n1     259:0    0 894,3G  0 disk
nvme1n1     259:1    0 894,3G  0 disk
```

Une fois le disque de stockage S0 migré, les données qui y sont sauvegardées peuvent être accédées en montant le système de fichiers :

[$ mount /dev/nvme1n1 /localckpt_restore/]

Les données sauvegardées récupérées sont alors transmises au nœud de calcul de rechange en charge de remplacer le nœud de calcul N1 en panne. Il peut s'agir du même nœud que le nœud N3 en charge de la récupération des données sauvegardées sur le disque de stockage S0.

Les données sauvegardées puis récupérées étant alors transmises au nœud de calcul désigné de rechange (« spare » en langue anglaise), il ne reste alors plus qu'à relancer l'application en changeant le nœud de calcul N1 en panne par ce dernier nœud de calcul de rechange dans la commande « mpirun » :

```
$ mpirun -host N2,N3 --mca pml ob1 --mca btl self,vader,tcp --mca coll ^hcoll,fca --mca oob_tcp_if_include enp12s0f0 --mca btl_tcp_if_include enp12s0f0 -n 2 hdf.exe
[ FTI  Information ] : Reading FTI configuration file (config.fti)...
[ FTI  Information ] : This is a restart. The execution ID is: 2016-11-09_17-09-10
FTI_Topo->nodeSize : 1 , FTI_Topo->nbProc : 2
FTI_Topo->nbNodes : 2 FTI_Topo->nbNodes : 2 FTI_Topo->groupSize : 2 , FTI_Topo->nodeSize 1
FTI_Topo->nodeSize : 1 , FTI_Topo->nbProc : 2
FTI_Topo->nbNodes : 2 FTI_Topo->nbNodes : 2 FTI_Topo->groupSize : 2 , FTI_Topo->nodeSize 1
[ FTI  Information ] : Recovering successfully from level 1.
[ FTI  Information ] : FTI has been initialized.
Local data size is  2048 x  1027 =   32. MB (   32).
Target precision :   0.00500
[ FTI  Information ] : Variable ID 0 to protect. Current ckpt. size per rank is 0.00MB.
[ FTI  Information ] : Variable ID 2 to protect. Current ckpt. size per rank is 16.05MB.
[ FTI  Information ] : Variable ID 1 to protect. Current ckpt. size per rank is 32.09MB.
Step :   500, error =   0.04888
```

Il peut être constaté que le calcul (qui détermine ici la valeur d'une erreur valant 0.04888) a donc bien repris à partir des sauvegardes locales de premier niveau L1.

Bien entendu, la présente invention n'est pas limitée aux exemples et au mode de réalisation décrits et représentés, mais elle est susceptible de nombreuses variantes accessibles à l'homme de l'art.

# REVENDICATIONS

1. Procédé de gestion de panne dans un réseau de nœuds (2), comprenant, pour chaque nœud considéré (2) de tout ou partie des nœuds (2) du réseau effectuant un même calcul :
   - d'abord, une étape de sauvegarde locale de l'état de ce nœud considéré (21), au niveau d'un support de stockage (31) de ce nœud considéré (21), le lien (6) entre ce support de stockage (31) et ce nœud considéré (21) pouvant être redirigé de ce support de stockage (31) vers un autre nœud (23),
   - ensuite, une étape de relance :
     - soit du nœud considéré (21) si celui-ci n'est pas défaillant, à partir de la sauvegarde locale de l'état de ce nœud considéré (21),
     - soit d'un nœud opérationnel (23) différent du nœud considéré (21), si le nœud considéré (21) est défaillant, à partir de la récupération de la sauvegarde locale de l'état de ce nœud considéré (21), en redirigeant ledit lien (6) entre le nœud considéré (21) et son support de stockage (31) de manière à relier ledit support de stockage (31) vers ledit nœud opérationnel (23),
   - les sauvegardes locales de ces nœuds considérés (2), utilisées pour les étapes de relance, sont cohérentes entre elles de sorte à correspondre à un même état de ce calcul.

2. Procédé de gestion de panne selon la revendication 1, caractérisé en ce que :
   - pour chaque nœud considéré non défaillant (22) :
     - il n'y a pas d'étape de récupération par un autre nœud (23), de la sauvegarde locale de l'état de ce nœud considéré non défaillant (22).

3. Procédé de gestion de panne selon la revendication 1 ou 2, caractérisé en ce que le nœud opérationnel (23) différent du nœud considéré (21) défaillant est un nœud de rechange (23).

4. Procédé de gestion de panne selon l'une quelconque des revendications précédentes, caractérisé en ce que ladite redirection dudit lien (6) entre le nœud considéré (21) et son support de stockage (31) de manière à relier ledit support de stockage (31) vers ledit nœud opérationnel (23), est réalisée par un changement d'aiguillage dans un commutateur (1) reliant plusieurs nœuds (21, 22, 23) à leurs supports de stockage (31, 32, 33).

5. Procédé de gestion de panne selon l'une quelconque des revendications précédentes, caractérisé en ce que toutes les étapes de relance de nœuds (2) sont synchronisées entre elles, de manière à relancer tous lesdits nœuds (2) dans un même état de calcul.

6. Procédé de gestion de panne selon l'une quelconque des revendications 1 à 5, caractérisé en ce que :
   - l'étape de récupération comprenant une sous-étape de migration, vers le nœud opérationnel (23) ou vers le nœud de rechange (23), du support de stockage (31) de la sauvegarde locale de l'état du nœud défaillant (21), qui rattache, ce support de stockage (31) de la sauvegarde locale de l'état du nœud défaillant (21), à un endroit prédéterminé de l'arborescence de fichiers de ce nœud opérationnel (23) ou de ce nœud de rechange (23), ce nœud opérationnel (23) ou ce nœud de rechange (23) venant ensuite lire, à cet endroit prédéterminé, le lien vers la sauvegarde locale de l'état du nœud défaillant (21), lors de l'étape de relance de ce nœud opérationnel (23) ou de ce nœud de rechange (23),

- aucune copie, de la sauvegarde locale de l'état du nœud défaillant (21), n'étant réalisée, au niveau d'un autre nœud (23).

7. Procédé de gestion de panne selon la revendication 6, caractérisé en ce que le nœud de rechange (23) et le nœud défaillant (21) qu'il remplace appartiennent tous les deux à la même lame de calcul.

8. Procédé de gestion de panne selon l'une quelconque des revendications 1 à 5, caractérisé en ce que l'étape de récupération comprend :
   - d'abord, une sous-étape de migration, vers un nœud intermédiaire (23), du support de stockage (31) de la sauvegarde locale de l'état du nœud défaillant (21), qui rattache, ce support de stockage (31) de la sauvegarde locale de l'état du nœud défaillant (21), à un endroit prédéterminé de l'arborescence de fichiers de ce nœud intermédiaire (23),
   - ensuite, une sous-étape de copie, de la sauvegarde locale de l'état du nœud défaillant (21), du support de stockage (31) rattaché au nœud intermédiaire (23) vers le support de stockage du nœud opérationnel ou du nœud de rechange, ce nœud opérationnel ou ce nœud de rechange venant ensuite lire son support de stockage, lors de son étape de relance.

9. Procédé de gestion de panne dans un réseau de nœuds, selon la revendication 8, caractérisé en ce que, pour tout ou partie des nœuds (2) du réseau effectuant un même calcul, le nœud de rechange (23) et le nœud défaillant (21) qu'il remplace appartiennent à des lames de calcul différentes.

10. Procédé de gestion de panne selon l'une quelconque des revendications précédentes, caractérisé en ce que toutes ces étapes sont réalisées pour tous les nœuds (2) du réseau effectuant un même calcul.

11. Procédé de gestion de panne selon l'une quelconque des revendications 6 à 10, caractérisé en ce que la sous-étape de migration change le rattachement du support de stockage (31) de la sauvegarde locale de l'état du nœud défaillant (21) en passant par l'intermédiaire d'un commutateur (1) auquel était rattaché le nœud défaillant (21) et son support de stockage (31) de la sauvegarde locale de l'état du nœud défaillant (21), mais sans passer par le nœud défaillant (21) lui-même.

12. Procédé de gestion de panne selon la revendication 11, caractérisé en ce que le changement de rattachement est réalisé par l'envoi d'une commande au commutateur (1), cette commande passant par l'un des nœuds (23) rattachés au commutateur (1) par un port de management (11).

13. Procédé de gestion de panne selon la revendication 11 ou 12, caractérisé en ce que ce commutateur (1) est un commutateur PCIe.

14. Procédé de gestion de panne selon l'une quelconque des revendications 11 à 13, caractérisé en ce que 3 à 10 nœuds (2) sont rattachés à un même commutateur (1).

15. Procédé de gestion de panne selon l'une quelconque des revendications précédentes, caractérisé en ce qu'il comprend aussi, pour tout ou partie des nœuds du réseau effectuant un même calcul, une étape de sauvegarde globale de l'ensemble de ces nœuds (2), réalisée moins souvent que l'ensemble des étapes de sauvegarde locales de ces nœuds (2).

16. Procédé de gestion de panne selon l'une quelconque des revendications 3 à 15, caractérisé en ce que, pour tout ou partie des nœuds (2) du réseau

effectuant un même calcul, le réseau comprend entre 1 et 5 nœuds de rechange (23) pour 100 nœuds (2) effectuant ledit même calcul.

17. Procédé de gestion de panne selon l'une quelconque des revendications précédentes, caractérisé en ce que, pour tout ou partie des nœuds (2) du réseau effectuant un même calcul, ce procédé de gestion de panne ne comprend aucun autre type de sauvegarde locale de l'état de ces nœuds (2).

18. Procédé de gestion de panne selon l'une quelconque des revendications précédentes, caractérisé en ce que, pour tout ou partie des nœuds (2) du réseau effectuant un même calcul, les supports de stockage (3) sont des mémoires flash.

19. Procédé de gestion de panne selon la revendication 18, caractérisé en ce que ces mémoires flash (3) sont des mémoires NVMe.

20. Partie d'un réseau de nœuds, comprenant :
    - un commutateur (1),
    - une pluralité de nœuds (2) d'un groupe de nœuds effectuant un même calcul dans ce réseau de nœuds,
    - plusieurs supports de stockage (3) respectivement rattachés à ces nœuds (2) par l'intermédiaire du commutateur (1),
    - ces supports de stockage (3) étant structurés et disposés pour sauvegarder localement l'état des nœuds (2) auxquels ils sont respectivement rattachés,
    - ce commutateur (1) étant structuré et disposé pour changer au moins un rattachement entre un nœud défaillant (21) et son support de stockage (31) lequel va alors devenir rattaché à un autre nœud (23) de ladite pluralité de nœuds (2).

21. Partie d'un réseau de nœuds selon la revendication 20, caractérisée en ce que ce commutateur (1) étant structuré et disposé pour changer au moins un rattachement entre un nœud défaillant (21) et son support de stockage (31), sur commande extérieure au commutateur (1).

22. Partie d'un réseau de nœuds selon la revendication 20 ou 21, caractérisée en ce que ladite pluralité de nœuds (2) comprend entre 3 et 10 nœuds.

23. Partie d'un réseau de nœuds selon l'une quelconque des revendications 20 à 22, caractérisée en ce que ce commutateur (1) est un commutateur PCIe.

24. Partie d'un réseau de nœuds selon l'une quelconque des revendications 20 à 23, caractérisée en ce que les supports de stockage (3) sont des mémoires flash, de préférence des mémoires NVMe.

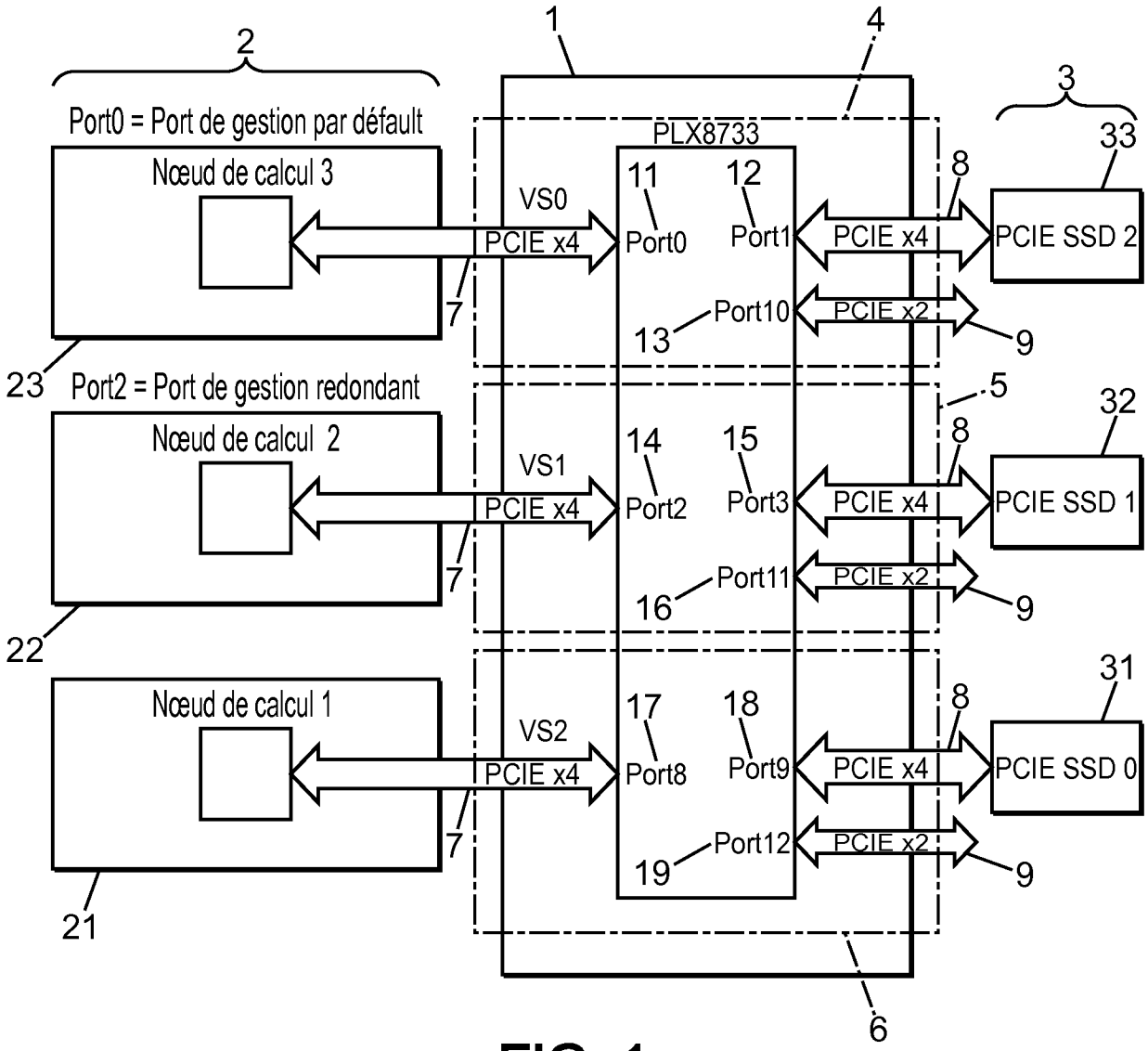
1
2
3
4
5
6
7
8
9
Port0 = Port de gestion par défaut
Nœud de calcul 3
Port2 = Port de gestion redondant
Nœud de calcul 2
Nœud de calcul 1
PLX8733
VS0
VS1
VS2
PCIE x4
PCIE x2
Port0
Port1
Port2
Port3
Port8
Port9
Port10
Port11
Port12
11
12
13
14
15
16
17
18
19
PCIE SSD 2
PCIE SSD 1
PCIE SSD 0
21
22
23
31
32
33

FIG. 1

# INTERNATIONAL SEARCH REPORT

International application No.
**PCT/FR2018/053535**

## A. CLASSIFICATION OF SUBJECT MATTER

***G06F 11/20***(2006.01)i; *G06F 11/14*(2006.01)n

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X | US 2014215264 A1 (ONO TAKATSUGU [JP] ET AL) 31 July 2014 (2014-07-31)<br>abstract; figure 2<br>paragraphs [0029] - [0031], [0035], [0041], [0045] - [0049], [0052], [0053], [0077] - [0079], [0103] - [0113], [0117]<br>paragraphs [0161], [0169] | 1-24 |
| X | US 2015355983 A1 (HARPER RICHARD E [US] ET AL) 10 December 2015 (2015-12-10)<br>abstract<br>paragraphs [0028] - [0034], [0038] - [0053] | 1-24 |
| A | US 2014281688 A1 (TIWARI PRAFULL [IN] ET AL) 18 September 2014 (2014-09-18)<br>abstract; figures 2-4<br>paragraphs [0005], [0013], [0017], [0018], [0020] | 1-24 |
| A | EP 0136560 A2 (NEC CORP [JP]) 10 April 1985 (1985-04-10)<br>abstract<br>page 6, line 20 - page 7, line 28<br>page 12, line 14 - line 22<br>page 13, line 20 - page 15, line 23 | 1-24 |

☑ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| **25 March 2019** | **03 April 2019** |

| Name and mailing address of the ISA/EP | Authorized officer |
|---|---|
| **European Patent Office**<br>**p.b. 5818, Patentlaan 2, 2280 HV Rijswijk**<br>**Netherlands**<br>Telephone No. **(+31-70)340-2040**<br>Facsimile No. **(+31-70)340-3016** | **Leuridan, Koen**<br>Telephone No. |

**INTERNATIONAL SEARCH REPORT**

International application No.
**PCT/FR2018/053535**

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| A | WO 2009140979 A1 (ERICSSON TELEFON AB L M [SE]; SPEKS OLIVIER HOLGER [DE]; WU JACKY [CN]) 26 November 2009 (2009-11-26)<br>the whole document | 1-24 |

**INTERNATIONAL SEARCH REPORT**
**Information on patent family members**

International application No.
**PCT/FR2018/053535**

| Patent document cited in search report | | | Publication date (day/month/year) | Patent family member(s) | | | Publication date (day/month/year) |
|---|---|---|---|---|---|---|---|
| US | 2014215264 | A1 | 31 July 2014 | JP | 6056509 | B2 | 11 January 2017 |
| | | | | JP | 2014146254 | A | 14 August 2014 |
| | | | | US | 2014215264 | A1 | 31 July 2014 |
| US | 2015355983 | A1 | 10 December 2015 | NONE | | | |
| US | 2014281688 | A1 | 18 September 2014 | NONE | | | |
| EP | 0136560 | A2 | 10 April 1985 | CA | 1208796 | A | 29 July 1986 |
| | | | | DE | 3483636 | D1 | 03 January 1991 |
| | | | | EP | 0136560 | A2 | 10 April 1985 |
| | | | | JP | S6054052 | A | 28 March 1985 |
| | | | | US | 4709325 | A | 24 November 1987 |
| WO | 2009140979 | A1 | 26 November 2009 | CN | 102037709 | A | 27 April 2011 |
| | | | | EP | 2283632 | A1 | 16 February 2011 |
| | | | | ES | 2649759 | T3 | 15 January 2018 |
| | | | | RU | 2010152239 | A | 27 June 2012 |
| | | | | US | 2011134749 | A1 | 09 June 2011 |
| | | | | US | 2014119171 | A1 | 01 May 2014 |
| | | | | WO | 2009140979 | A1 | 26 November 2009 |

# RAPPORT DE RECHERCHE INTERNATIONALE

Demande internationale n°
PCT/FR2018/053535

A. CLASSEMENT DE L'OBJET DE LA DEMANDE
INV. G06F11/20
ADD. G06F11/14

Selon la classification internationale des brevets (CIB) ou à la fois selon la classification nationale et la CIB

B. DOMAINES SUR LESQUELS LA RECHERCHE A PORTE

Documentation minimale consultée (système de classification suivi des symboles de classement)
G06F

Documentation consultée autre que la documentation minimale dans la mesure où ces documents relèvent des domaines sur lesquels a porté la recherche

Base de données électronique consultée au cours de la recherche internationale (nom de la base de données, et si cela est réalisable, termes de recherche utilisés)
EPO-Internal

C. DOCUMENTS CONSIDERES COMME PERTINENTS

| Catégorie* | Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents | no. des revendications visées |
|---|---|---|
| X | US 2014/215264 A1 (ONO TAKATSUGU [JP] ET AL) 31 juillet 2014 (2014-07-31)<br>abrégé; figure 2<br>alinéas [0029] - [0031], [0035], [0041], [0045] - [0049], [0052], [0053], [0077] - [0079], [0103] - [0113], [0117]<br>alinéas [0161], [0169]<br>----- | 1-24 |
| X | US 2015/355983 A1 (HARPER RICHARD E [US] ET AL) 10 décembre 2015 (2015-12-10)<br>abrégé<br>alinéas [0028] - [0034], [0038] - [0053]<br>----- | 1-24 |
| A | US 2014/281688 A1 (TIWARI PRAFULL [IN] ET AL) 18 septembre 2014 (2014-09-18)<br>abrégé; figures 2-4<br>alinéas [0005], [0013], [0017], [0018], [0020]<br>----- | 1-24 |

-/--

[X] Voir la suite du cadre C pour la fin de la liste des documents

[X] Les documents de familles de brevets sont indiqués en annexe

* Catégories spéciales de documents cités:

"A" document définissant l'état général de la technique, non considéré comme particulièrement pertinent

"E" document antérieur, mais publié à la date de dépôt international ou après cette date

"L" document pouvant jeter un doute sur une revendication de priorité ou cité pour déterminer la date de publication d'une autre citation ou pour une raison spéciale (telle qu'indiquée)

"O" document se référant à une divulgation orale, à un usage, à une exposition ou tous autres moyens

"P" document publié avant la date de dépôt international, mais postérieurement à la date de priorité revendiquée

"T" document ultérieur publié après la date de dépôt international ou la date de priorité et n'appartenant pas à l'état de la technique pertinent, mais cité pour comprendre le principe ou la théorie constituant la base de l'invention

"X" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme nouvelle ou comme impliquant une activité inventive par rapport au document considéré isolément

"Y" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme impliquant une activité inventive lorsque le document est associé à un ou plusieurs autres documents de même nature, cette combinaison étant évidente pour une personne du métier

"&" document qui fait partie de la même famille de brevets

| Date à laquelle la recherche internationale a été effectivement achevée | Date d'expédition du présent rapport de recherche internationale |
|---|---|
| 25 mars 2019 | 03/04/2019 |

| Nom et adresse postale de l'administration chargée de la recherche internationale | Fonctionnaire autorisé |
|---|---|
| Office Européen des Brevets, P.B. 5818 Patentlaan 2<br>NL - 2280 HV Rijswijk<br>Tel. (+31-70) 340-2040,<br>Fax: (+31-70) 340-3016 | Leuridan, Koen |

1

# RAPPORT DE RECHERCHE INTERNATIONALE

Demande internationale n°
PCT/FR2018/053535

C(suite). DOCUMENTS CONSIDERES COMME PERTINENTS

| Catégorie* | Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents | no. des revendications visées |
|---|---|---|
| A | EP 0 136 560 A2 (NEC CORP [JP])<br>10 avril 1985 (1985-04-10)<br>abrégé<br>page 6, ligne 20 - page 7, ligne 28<br>page 12, ligne 14 - ligne 22<br>page 13, ligne 20 - page 15, ligne 23<br>----- | 1-24 |
| A | WO 2009/140979 A1 (ERICSSON TELEFON AB L M [SE]; SPEKS OLIVIER HOLGER [DE]; WU JACKY [CN]) 26 novembre 2009 (2009-11-26)<br>le document en entier<br>----- | 1-24 |

1

# RAPPORT DE RECHERCHE INTERNATIONALE

**Renseignements relatifs aux membres de familles de brevets**

**Demande internationale n°**
PCT/FR2018/053535

| Document brevet cité au rapport de recherche | Date de publication | Membre(s) de la famille de brevet(s) | Date de publication |
|---|---|---|---|
| US 2014215264 A1 | 31-07-2014 | JP 6056509 B2 | 11-01-2017 |
| | | JP 2014146254 A | 14-08-2014 |
| | | US 2014215264 A1 | 31-07-2014 |
| US 2015355983 A1 | 10-12-2015 | AUCUN | |
| US 2014281688 A1 | 18-09-2014 | AUCUN | |
| EP 0136560 A2 | 10-04-1985 | CA 1208796 A | 29-07-1986 |
| | | DE 3483636 D1 | 03-01-1991 |
| | | EP 0136560 A2 | 10-04-1985 |
| | | JP S6054052 A | 28-03-1985 |
| | | US 4709325 A | 24-11-1987 |
| WO 2009140979 A1 | 26-11-2009 | CN 102037709 A | 27-04-2011 |
| | | EP 2283632 A1 | 16-02-2011 |
| | | ES 2649759 T3 | 15-01-2018 |
| | | RU 2010152239 A | 27-06-2012 |
| | | US 2011134749 A1 | 09-06-2011 |
| | | US 2014119171 A1 | 01-05-2014 |
| | | WO 2009140979 A1 | 26-11-2009 |