L'atout de l'informatique d'entreprise moderne

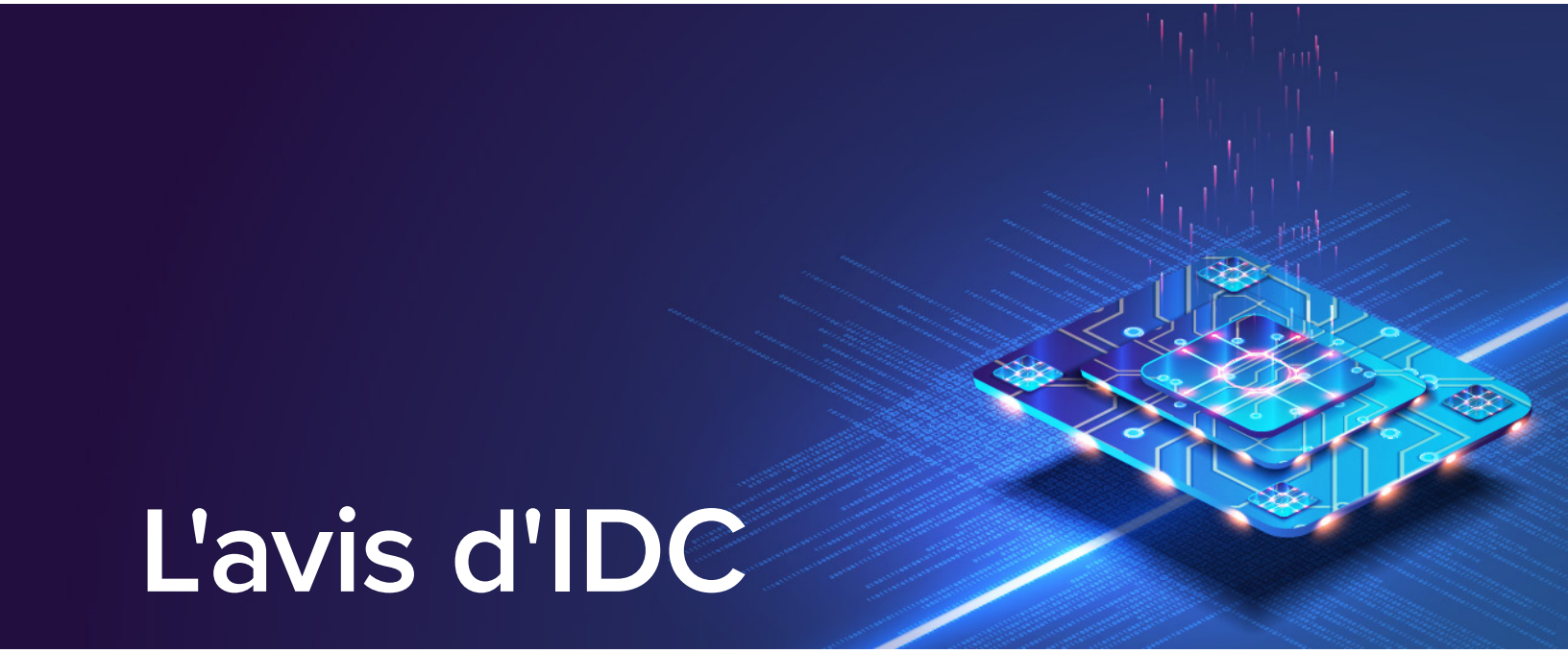
Auteur :

Peter Rutten
Directeur de recherche, Infrastructure Systems,
Platforms and Technologies Group, Performance
Intensive Computing Solutions Global Research Lead, IDC

Septembre 2021 | IDC Doc. #US48197921

Navigation dans ce livre blanc

Cliquez sur les titres ou sur les numéros de page pour accéder à chaque section.



L'avis d'IDC

Le paysage informatique actuel peut sembler énigmatique. Contraintes d'opérer leur transformation numérique et de satisfaire les besoins de clients extrêmement exigeants, les entreprises s'efforcent aujourd'hui de mener à bien une mission quasi-impossible.

- Les marchés peuvent changer du tout au tout, passant de périodes de pics à des périodes creuses, et cette volatilité ne peut pas être considérée comme exceptionnelle. Elle constitue la norme actuelle.
- Pour satisfaire aux flux et reflux des charges de travail liés à la demande, les systèmes doivent s'adapter parfaitement et dynamiquement sans qu'il soit nécessaire de créer un datacenter important, coûteux et énergivore pour les seules périodes de pics. La durabilité n'est plus un simple gadget marketing.
- La complexité de ces marchés ne peut plus être analysée et exploitée via l'expérience et l'intelligence humaines seules. Une grande partie de l'intelligence doit désormais être artificielle, fonctionner en temps réel et jongler avec d'innombrables variables tout en intégrant d'énormes quantités de données. L'intelligence artificielle (IA) va être de plus en plus omniprésente ; or, elle nécessite des dispositifs matériels dédiés.
- Devant l'exigence de disponibilité perpétuelle, les charges de travail sur lesquelles repose l'entreprise numérique ne peuvent être ni ralenties ni entravées, encore moins être indisponibles. Dans notre monde perpétuellement connecté, toute indisponibilité peut être catastrophique.
- Le tout numérique et connecté lié à l'avènement des entreprises numériques est exposé à de nouveaux types d'attaques qui peuvent le compromettre. Des communautés entières de personnes mal intentionnées se sont regroupées dans un monde parallèle souterrain qui mène une guerre permanente contre les entreprises du monde entier par le biais d'un vaste arsenal d'outils et de stratégies de cyberattaque. Disposer d'une sécurité complète et inviolable est donc désormais incontournable.

Pour qu'une plateforme de traitement puisse faire office de moteur de l'entreprise numérique, elle doit donc être parfaitement sécurisée, fiable, évolutive, durable, être en mesure de s'intégrer au cloud dans le cadre d'une approche hybride, et être conçue pour l'IA. Ce livre blanc traite en détail ces considérations du point de vue de l'infrastructure et du déploiement. En outre, il examine comment se positionnent le nouveau processeur IBM Power 10 et la nouvelle plateforme d'entreprise IBM Power à leur égard.

Aperçu de la situation

Pour IDC, une entreprise numérique qui veut réussir dans l'environnement complexe et multiforme actuel, doit prendre en compte ces considérations critiques :

- **La sécurité : une exigence incontournable**
- **L'obligation de fiabilité**
- **Évolutivité et durabilité**
- **L'infrastructure informatique hybride adéquate (cloud hybride et applications natives cloud)**
- **L'importance de l'IA et où l'exécute**

Dans les sections suivantes, chacune de ces considérations est examinée en détail.

La sécurité : une exigence incontournable

La sécurité est devenue l'exigence le plus importante de l'entreprise numérique. Lorsqu'IDC interroge des entreprises sur leurs priorités, la sécurité figure invariablement dans le groupe de tête. D'ailleurs, si on leur demande, par exemple, d'indiquer les principaux éléments de l'infrastructure IA qui ne sont pas optimaux dans les offres de leurs fournisseurs de serveurs et de stockage, la sécurité obtient le score le plus élevé, puisque 30 % d'entre-elles se disent insatisfaites des fonctions de sécurité.[1]

Par conséquent, nombre d'entre-elles n'autorisent pas l'utilisation des dispositifs de stockage qui contiennent les données de leurs charges de travail IA par d'autres charges de travail. Le motif le plus souvent invoqué (45 %) est la sécurité et la confidentialité des données. En outre, les résultats de l'enquête d'IDC montrent que la sécurité est une préoccupation majeure dans les infrastructures de cloud public en tant que service (IaaS), 37 % des entreprises déclarant que la sécurité est leur principal enjeu dans ces déploiements.[2] De plus en plus, les entreprises injectent de l'IA, majoritairement dans leurs charges de travail de sécurité, afin de pouvoir mieux prévoir les violations et y réagir.

Elles se consacrent et investissent majoritairement dans la sécurité des piles d'applications et réseau. Un grand nombre d'attaques, cependant, sont de bas niveau et centrées sur le matériel. Elles tirent souvent parti des vulnérabilités des processeurs et/ou du microcode. Ces attaques sont sophistiquées et difficiles à détecter.

IDC constate donc que les entreprises s'intéressent de plus en plus à l'« informatique confidentielle » pour leurs plateformes métier critiques. L'informatique confidentielle permet l'isolement des données sensibles dans un sous-système de processeurs désigné et protégé (parfois appelé « enclave de processeurs sécurisée ») en vue de leur traitement. Aujourd'hui, les données sont souvent chiffrées au repos, quand elles sont stockées et en transit sur le réseau, mais pas lorsqu'elles sont utilisées en mémoire.

[1] Source : IDC AI Infrastructure View 2021

[2] Source : IDC IaaSView 2020

La capacité de protection de données et de code en mémoire est limitée dans de nombreuses plateformes informatiques. Pourtant, les entreprises qui traitent des données sensibles telles que des informations personnelles, des données financières ou des informations sur la santé, ont besoin d'atténuer les menaces qui visent les applications ou les données dans la mémoire système.

Dans le cadre de l'informatique confidentielle, le contenu du sous-système, qui peut être chiffré au niveau matériel, n'est accessible qu'à du code autorisé dans un programme. Il est inaccessible à tout ce qui lui est extérieur, y compris à d'autres codes, systèmes ou opérateurs. Les entités non autorisées ne peuvent pas voir ou perturber les données ni le processus d'exécution du code autorisé. Une solution complète d'informatique confidentielle sécurise les données en cours d'utilisation et au repos via le chiffrement du contenu dans une mémoire système volatile ou non volatile et dans des magasins de données persistantes, sur des supports flash ou des disques durs.

Les infrastructures d'informatique confidentielle modernes, en particulier celles implémentées dans des environnements mutualisés partagés, utilisent des coprocesseurs discrets pour décharger des opérations de processeurs privilégiés qui peuvent être compromises par des vulnérabilités d'exécution de code de bas niveau. Ce n'est pas encore une approche courante, mais pour les charges de travail de base des entreprises, elle est très prometteuse. Dans l'intervalle, les entreprises exploitent diverses stratégies de sécurité simultanément, tant au niveau du matériel que des logiciels.

L'obligation de fiabilité

Si les stratégies de sécurité sont d'une importance capitale pour protéger les données, les applications et le matériel contre les attaques, un autre aspect crucial de l'entreprise numérique est la fiabilité absolue de l'environnement informatique, y compris de l'infrastructure. La haute disponibilité n'est pas un nouveau concept et les entreprises peuvent sélectionner des plateformes de traitement offrant un taux de disponibilité allant jusqu'à 99,999 % et des plateformes de stockage où il est égal à 99,99999 %. Mais ces chiffres ne peuvent être atteints qu'avec le matériel, les logiciels et les politiques adéquats. IDC n'a attribué le niveau de disponibilité 4 (AL4)[3] qu'à neuf plateformes de serveurs de six fournisseurs sur le marché des serveurs ; il s'agit du niveau le plus élevé, qui représente une tolérance totale aux pannes.

- L'étude IDC[4] montre que les trois principales causes de l'indisponibilité des applications sont la défaillance du réseau (16,2 %), la défaillance des serveurs (15,5 %), et les logiciels malveillants (10,3 %). Les causes les plus courantes de défaillance des serveurs sont la surcharge de mémoire (DRAM) ou d'UC, la défaillance ou l'altération de mémoire.
- Les volumes de transactions augmentent de façon spectaculaire et les entreprises ont besoin de vitesses de transaction toujours plus élevées pour satisfaire leurs clients.
- Les charges de travail critiques sont de plus en plus nombreuses, et les fonctions de support métier qui pouvaient auparavant être exécutées sur un niveau à faible disponibilité, par exemple, via la virtualisation ou le regroupement, sont de plus en plus considérées comme critiques pour l'entreprise.
- Le coût de l'indisponibilité augmente en même temps que la dépendance des entreprises vis-à-vis de leur infrastructure pour leurs opérations quotidiennes. L'étude IDC montre que pour 20,7 % des entreprises, le coût de l'indisponibilité est de 5000 à 10 000 dollars par heure ; pour 18,4 %, il est de 10 000 à 25 000 dollars par heure; pour 17 %, il est de 25 000 à 100 000 dollars par heure; et pour certaines entreprises (1,4 %), il est de 500 000 dollars par heure.
- Avec la fin des « heures de bureau normales », c'est-à-dire avec la possibilité pour les clients d'utiliser les applications d'entreprise en permanence, la pression sur l'infrastructure qui prend en charge ces applications s'est considérablement accentuée car les interruptions imprévues sont désormais très peu, voire pas du tout, tolérées.
- Ni les entreprises ni les clients ne tolèrent de pannes, de retards, de perte de données ni d'altération ; toute violation ou erreur peut avoir des conséquences catastrophiques pour la réputation d'une entreprise.

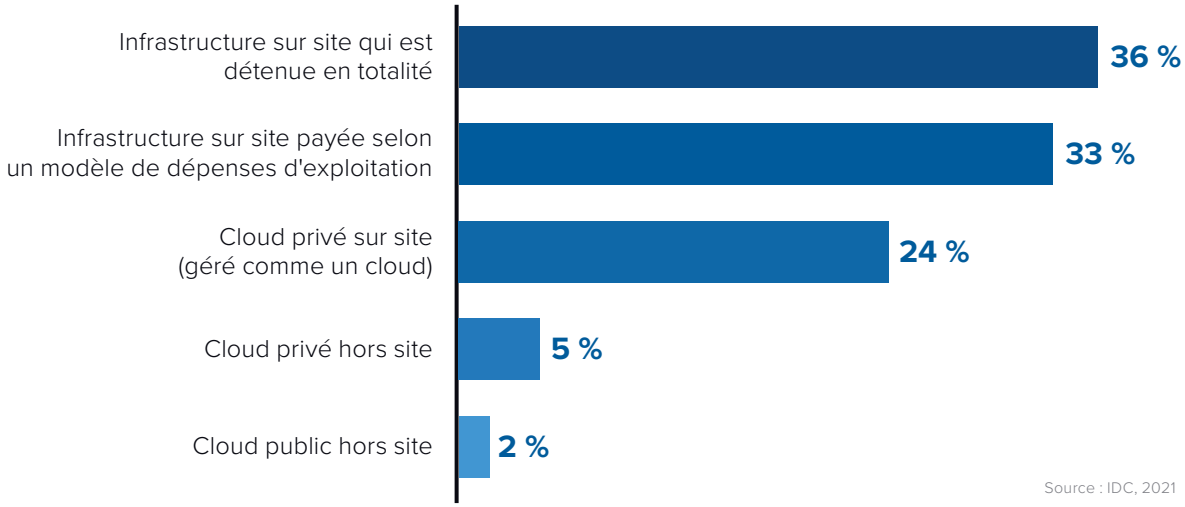
[3] Source : IDC Worldwide AL4 Server Market Shares, 2019: *Fault-Tolerant Systems Become Digital Transformation Platforms*

[4] Source : IDC *Server Storage Infrastructure Availability Survey,* 2018

- Étant donné que les entreprises interagissent numériquement plus fréquemment et de façons très diverses avec les consommateurs, les citoyens et les autres entreprises, la conformité avec les réglementations nationales et internationales sur la disponibilité, la sécurité et la confidentialité des données est d'une importance capitale.
- Même si la disponibilité et la sécurité dans le cloud public se sont considérablement améliorées, la véritable tolérance aux pannes continue d'être considérée comme une fonctionnalité sur site ou du cloud hybride, et non comme celle d'un cloud public (voir **Figure 1**).

FIGURE 1

Infrastructure qui héberge le niveau à plus haute disponibilité

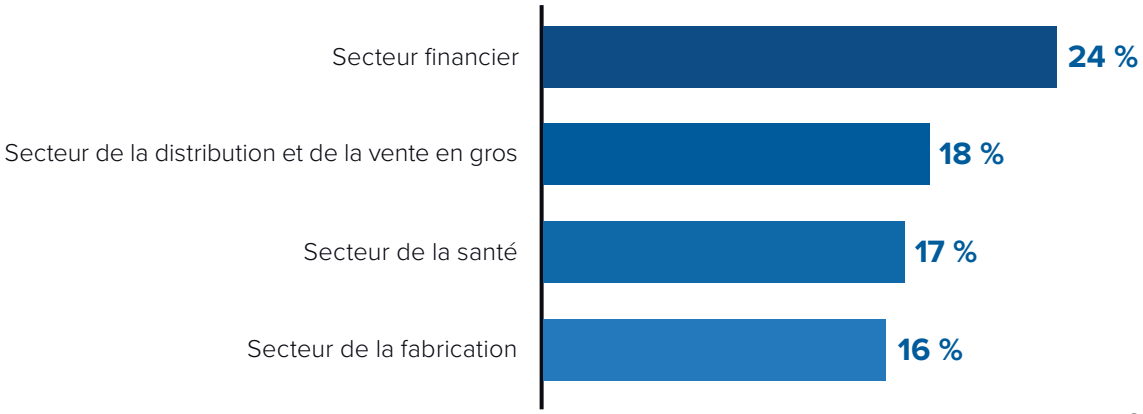


Source : IDC, 2021

Le pourcentage des systèmes qui doivent être hautement disponibles augmente en conséquence. Dans tous les secteurs d'activité, pour plus de 60 % des entreprises, 21 à 30 % de leurs serveurs se trouvent au niveau à plus haute disponibilité. La **figure 2** montre le pourcentage de systèmes qui doivent être hautement disponibles dans divers secteurs.

FIGURE 2

Pourcentage de systèmes qui doivent être hautement disponibles, par secteur



Source : IDC, 2021

Les plateformes AL4 majeures ont progressé au point d'être totalement intégrées au datacenter ; elles ne se contentent pas de participer à la transformation numérique : elles la dirigent. Ces systèmes traitent les données les plus critiques et les plus précieuses de nombreuses entreprises, dont les volumes sont souvent plus importants que ceux des autres types de données. Elles doivent libérer ces données et en tirer parti pour devenir des entreprises numériques.

Le besoin d'évolutivité et de durabilité

Les entreprises ont besoin de faire évoluer les charges de travail qui traitent des volumes toujours croissants de données dans des environnements informatiques en constant développement.

En parallèle, elles doivent pouvoir s'adapter à la hausse ou à la baisse selon les niveaux de la demande parfois imprévisibles, qui peuvent atteindre des pics extrêmement élevés. Pour ce faire, elles requièrent des datacenters plus importants, davantage d'équipements, de renouvellements de ceux-ci, ainsi qu'une énergie accrue pour les faire fonctionner, tout en les refroidissant.

Les charges de travail d'IA sont celles qui connaissent la croissance la plus rapide, consomment des données et font l'objet d'investissements informatiques effectués par les entreprises. 21 % des entreprises disent investir dans des technologies informatiques assurant le traitement parallèle nécessaire pour la formation et l'inférence sur les réseaux d'apprentissage profond en IA. 9 % supplémentaires des entreprises disent qu'elles prévoient de le faire en 2021. En outre, 46 % des entreprises investissent dans les technologies d'accélération de charge de travail telles que les GPU, les FPGA et les ASIC ; 7 % supplémentaires prévoient d'investir en 2021.[5] Ces derniers, en particulier, ont donné lieu à des problèmes au niveau des datacenters en ce qui concerne les exigences énergétiques et le refroidissement. Le scénario d'utilisation le plus courant pour l'accélération est l'inférence en apprentissage profond en IA (mettre en production un modèle d'IA développé avec un réseau de neurones profond [DNN]). 38 % des entreprises utilisent l'accélération pour l'inférence en IA, alors que seulement 27 % l'utilisent pour la formation d'un DNN.[6] Cette tendance, à savoir que les investissements en traitement de l'inférence en IA commencent à dépasser ceux de la formation en IA, était attendue. De plus, l'IA n'est pas la seule charge de travail qui motive les investissements dans l'accélération via de tels coprocesseurs. L'analyse de données, le HPC, la modélisation financière, la cybersécurité et la détection des fraudes, ainsi que les négociations financières, sont d'autres exemples de charges de travail qui s'exécutent de plus en plus sur des GPU, des FPGA ou des ASIC, et une majorité d'entreprises les exécutent sur site.

21 % des entreprises disent investir dans des technologies informatiques qui permettent d'effectuer le traitement parallèle nécessaire pour la formation et l'obtention d'inférences sur les réseaux d'apprentissage profond en IA.

Il existe toutefois une problématique majeure ; la plupart des datacenters ne sont pas équipés pour supporter plusieurs racks de nœuds de traitement accéléré, à savoir, qu'ils ne disposent pas de la puissance qu'ils nécessitent, ni de la capacité de dissipation de la chaleur qu'ils génèrent, qui est beaucoup plus importante qu'avec des racks de serveurs non accélérés. Selon le ministère américain de l'énergie (2020), les datacenters sont l'un des types de bâtiments les plus énergivores ; en effet, ils consomment 10 à 50 fois plus d'énergie par espace au sol qu'un bâtiment de bureaux commerciaux classique. IDC a pu déterminer que, en moyenne, 17,6 % du budget d'exploitation d'un datacenter est consacré à l'électricité, plus que tout autre article budgétaire. Aux États-Unis, les datacenters représentent 2 % de l'électricité totale utilisée dans le secteur commercial.

Toutefois, de nombreuses entreprises, notamment dans le secteur technologique, tentent d'améliorer leur empreinte carbone. Les sociétés technologiques sont en tête de la liste EPA (Environmental Protection Agency) des entreprises vertes et IDC a constaté d'énormes investissements dans les énergies renouvelables du secteur technologique, ainsi que des investissements dans du matériel et des logiciels plus respectueux de l'environnement qui permettent de réduire la consommation énergétique. IDC a constaté que ces derniers ont permis de réduire la consommation énergétique de 26 % en moyenne.

[5] Source : IDC IT Infrastructure Plans for 2021 Survey, 2020

[6] Source : IDC IT Infrastructure for Compute Survey, 2021

De nombreuses entreprises ont pris exemple sur les fournisseurs de services cloud concernant une approche plus durable de leur informatique, notamment en réutilisant et en recyclant leur équipement ; 33 % des personnes interrogées dans le cadre d'une enquête IDC[7] ont déclaré que cela contribuait à atteindre une plus grande durabilité. La réutilisation et le recyclage de l'équipement peuvent en effet contribuer de manière significative à l'empreinte carbone globale d'un datacenter. Mettre à niveau certains composants d'un serveur peut être requis, mais le nombre de nouveaux composants indispensables d'une génération de serveurs à une autre n'est pas supérieur à celui des composants qui pourraient simplement être conservés et être réutilisés.

Les entreprises prennent de plus en plus conscience de l'intérêt de la réutilisation pour réduire leur empreinte environnementale ; IDC prévoit que d'ici 2025, 90 % des entreprises du G2000 vont exiger des matériaux réutilisables dans les chaînes d'approvisionnement de matériel, des objectifs en neutralité carbone pour les installations des fournisseurs, et une utilisation énergétique moindre préalable à la conclusion de contrats.[8] Ces mesures contribuent également à réduire les coûts pour les entreprises, qu'il s'agisse d'une utilisation énergétique moindre ou d'investissements en matériel réduits.

L'infrastructure informatique hybride adaptée

Un penchant pour le cloud hybride

Aujourd'hui, 54 % des applications d'entreprises sont encore déployées sur site.[9] Selon IDC, ce pourcentage ne va pas baisser de manière significative ; les entreprises indiquent que dans deux ans, elles pensent toujours exécuter 52 % de leurs applications sur site. Parmi ces applications sur site, 56 % s'exécutent en tant que cloud privé, un chiffre qui devrait passer à 60 % dans deux ans. Quant à savoir si le cloud privé répond à leurs objectifs, 61 % des entreprises affirment qu'il répond non seulement à leurs attentes mais les dépasse.

Nombre de ces applications, en particulier les applications métier critiques, présentent des interdépendances complexes. En moyenne, les entreprises déclarent que 49 % de leurs applications métier ont des dépendances et 27 % ont des interdépendances complexes. Aujourd'hui, seules 18 % des applications sont considérées comme « natives cloud », à savoir qu'il s'agit de microservices modulaires et distincts qui représentent des suites de services pouvant être déployés indépendamment. En revanche, 32 % des applications continuent à être monolithiques. Mais cela va changer très rapidement. Les entreprises indiquent que dans deux ans, seules 21 % des applications métier critiques seront monolithiques, tandis que 44 % seront natives cloud.

Parallèlement, les entreprises pensent tirer parti de différents déploiements cloud sur et hors site, ce qui est souvent désigné comme une approche de cloud « hybride ». IDC considère que ce scénario est en plein essor. La **figure 3** montre que la combinaison cloud la plus courante consiste à disposer de plusieurs clouds pour faire migrer des charges de travail et des données entre eux. En ce qui concerne le scénario « cloud privé/cloud public », environ 40 % des entreprises indiquent que ces deux déploiements interopèrent en leur sein ; en d'autres termes, ils constituent un cloud hybride plus ou moins intégré.

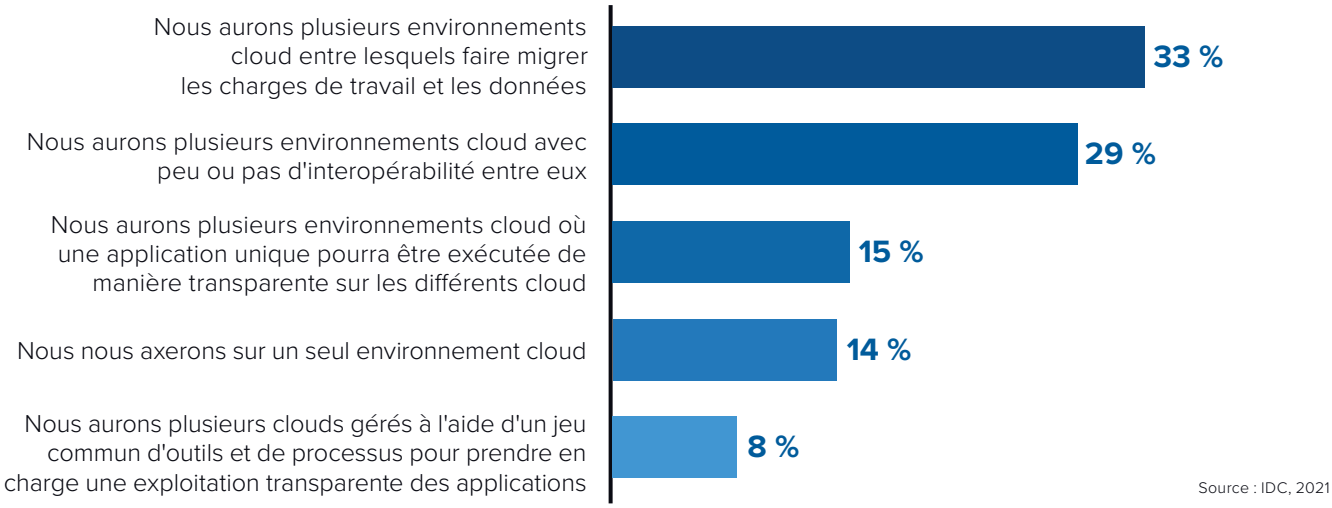
Notons que pour la partie sur site d'un cloud hybride, les entreprises souhaitent très majoritairement (84 %) passer d'un modèle basé sur les dépenses d'investissement à un modèle basé sur les dépenses d'exploitation. Actuellement, 42 % des budgets informatiques des entreprises sont financés par une approche basée sur les dépenses d'exploitation ; il y a trois ans, ce chiffre était de 36 %.

Notons que pour la partie sur site d'un cloud hybride, les entreprises souhaitent très majoritairement (84 %) passer d'un modèle basé sur les dépenses d'investissement à un modèle basé sur les dépenses d'exploitation.

[7] Source : IDC 2021 Datacenter Operational Survey
[8] Source : IDC Worldwide Future of Digital Infrastructure 2021 Predictions
[9] Source : IDC 1Q21 Cloud Pulse Survey, May 2021

FIGURE 3

Utilisation d'environnements cloud sur et hors site

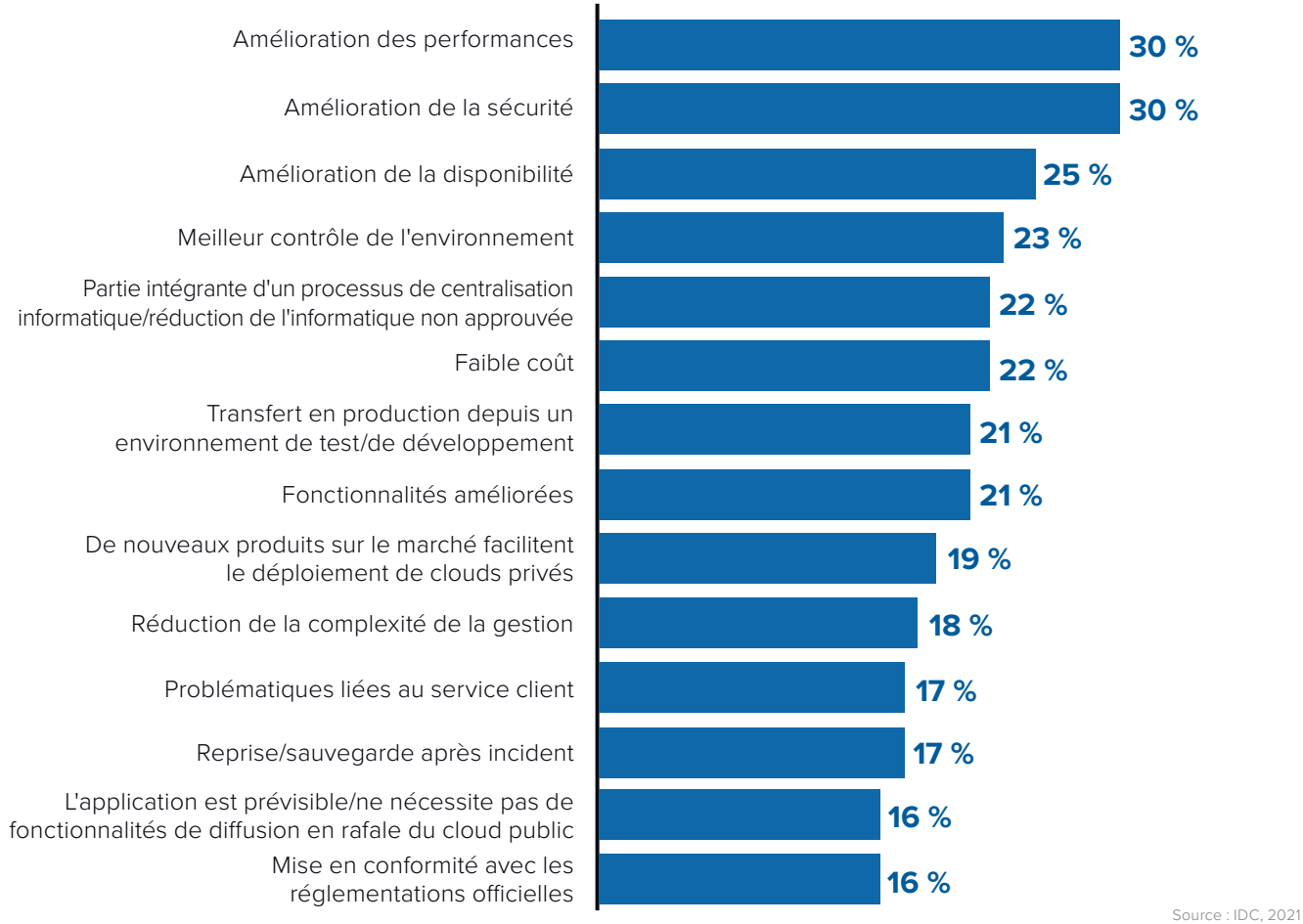


Source : IDC, 2021

Si le cloud hybride est de plus en plus répandu, le rapatriement d'un cloud public vers un cloud privé est également très courant : 66 % des entreprises déclarent transférer des applications vers leurs environnements de cloud privé ou non cloud pour diverses raisons, les performances, la sécurité et la disponibilité étant les trois principales (voir **Figure 4**).

FIGURE 4

Motifs du transfert d'applications du type IaaS vers un cloud privé ou un environnement non cloud



Source : IDC, 2021

Cloud hybride et applications natives cloud

Un cloud hybride correctement conçu est une plateforme idéale pour développer et exécuter des applications natives cloud, que de plus en plus d'entreprises considèrent comme une fonctionnalité importante pour leur transformation numérique. IDC a révélé qu'une majorité d'entreprises considèrent l'implémentation de fonctionnalités diverses comme étant « importante » à « extrêmement importante » pour répondre à leurs besoins métier dans le cadre de leur investissement dans une stratégie cloud permettant le développement et l'exécution d'applications natives cloud. Ces fonctionnalités sont les suivantes :

- Amélioration des performances, de la disponibilité, de la portabilité et de la gestion des applications.
- Amélioration de l'intégration de données, de l'orchestration, de l'observabilité, de la gestion d'API et des AIOps sur des environnements cloud
- Accélération des cycles de développement et de la mise sur le marché grâce au CI/CD (développement et déploiement continus) et à l'automatisation
- Politiques de sécurité globales, gestion des risques, stratégies de reprise après incident et conformité aux réglementations
- Modèle basé sur les dépenses d'exploitation plutôt que sur les dépenses d'investissement, incluant des fonctionnalités de rétrofacturation
- Optimisation de la productivité, de l'efficacité et des compétences du personnel.

Les entreprises qui veulent accroître leurs investissements dans un cloud hybride doivent s'assurer qu'elles disposent de toutes ces fonctionnalités afin d'obtenir le retour sur investissement qu'elles anticipent.

L'importance de l'IA et où l'exécuter

IDC prévoit que le marché mondial des plateformes de serveurs d'IA atteindra 27 milliards de dollars d'ici 2025.[10]

En effet, on prévoit une adoption accrue des technologies conversationnelles, du traitement automatique du langage naturel, de l'analyse d'images et de vidéos, de l'apprentissage profond, de l'apprentissage automatique, de la génération d'hypothèses et de l'analyse prédictive. Les plateformes de serveurs d'IA constitueront 21 % du marché mondial des serveurs d'ici 2025.

Dans une section précédente, nous avons indiqué la nécessité croissante de disposer de coprocesseurs afin d'exécuter les charges de travail de formation et d'inférence en IA. Étant donné que le cloud privé sur site est le premier scénario de déploiement pour l'IA, suivi par un environnement non cloud sur site, les entreprises investissent massivement en GPU, FPGA et ASIC. Pour la formation de l'IA, ces investissements sont plus ou moins inévitables ; former un algorithme DNN ne peut tout simplement pas être effectué sur un processeur hôte. Mais pour l'inférence en IA, il existe de nombreux modèles d'IA qui s'exécuteront parfaitement sur un processeur hôte avancé ou sur un processeur hôte doté d'un processeur IA spécialisé intégré. Ces scénarios présentent un net avantage en termes de coût pour les entreprises, étant donné que l'ajout de quelques GPU à un serveur peut rapidement doubler le prix total.

Étant donné que le cloud privé sur site est le premier scénario de déploiement pour l'IA, suivi par un environnement non cloud sur site, les entreprises investissent massivement.

Pourquoi les entreprises continuent-elles à donner la préférence à l'exécution de leurs applications d'IA sur site ? Pourquoi ne pas les exécuter dans le cloud, par exemple, pour ne plus avoir de dépenses d'investissement ? Bien sûr, une partie de la formation de l'IA est effectuée dans des clouds publics sur les plateformes d'IA de fournisseurs, et une fois développés, ces modèles restent parfois dans le cloud comme charges de travail de production.

[10] IDC Worldwide AI Server Forecast, 2021–2025, juillet 2021

Le choix du cloud ou d'un environnement sur site repose sur les données ; il est lié aux questions ci-après :

Quelles sont les données nécessaires pour développer le modèle ?
S'il s'agit de données provenant d'applications d'entreprise clés telles que des données transactionnelles, il est préférable qu'elles restent sur la plateforme transactionnelle, notamment pour des raisons de latence.

Ces données sont-elles sensibles ?
Si les données sont sensibles, ce qui signifie qu'elles doivent être extrêmement bien protégées, il n'est pas souhaitable de les déplacer vers le cloud, que ce soit pour la formation ou l'inférence.

Quel est le cadre réglementaire concernant les données ?
Certaines données ne peuvent légalement pas être déplacées vers un cloud public ; c'est très souvent le cas des données clés des entreprises. Les entreprises sont soumises à toutes sortes de réglementations, qu'il s'agisse de réglementations nationales sur la protection des données, du RGPD, des réglementations sectorielles telles que la loi HIPAA, des réglementations ISO ou du California Consumer Protection Act.

Que peut-on et ne peut-on pas faire avec les données pour rester en conformité ?
Lorsque les données commencent à être déplacées, il devient difficile de garantir que l'entreprise reste conforme.

Quel est le volume des données ?
Plus le nombre de données requises pour la formation ou utilisées pour l'inférence par le modèle d'IA est élevé, surtout si cette inférence est en quasi-temps réel, plus il devient difficile de le faire dans le cloud.

Quel est le degré d'intégration des applications qui exploitent les données ?
La plateforme qui exécute les transactions dispose très probablement de multiples applications profondément intégrées à la base de données pour assurer l'analyse et d'autres fonctions, ce qui complexifie leur transfert vers le cloud.

Quel est le coût du stockage des données ?
Le stockage de volumes importants dans le cloud peut rapidement dépasser les dépenses d'investissement qui seraient nécessaires pour les stocker sur site.

Toutes ces considérations amènent de nombreuses entreprises à conserver leurs charges de travail de formation et d'inférence d'IA sur site. Elles peuvent continuer à former sur un environnement de traitement distinct dans le datacenter derrière leur pare-feu, mais elles replacent ensuite le modèle formé sur la plateforme qui exécute les applications clés pour inférence. Si la plateforme permet une inférence robuste, les entreprises peuvent utiliser l'IA sur des données clés qui étaient auparavant hors limites.

IBM Power10 et IBM Power E1080

Pour réussir leur transformation en entreprise numérique, les entreprises ont besoin de plateformes de traitement pouvant absorber la volatilité du marché, qui sont sécurisées sans compromis, s'adaptent facilement tout en réduisant leur empreinte physique et carbone, offrent les plus hauts niveaux de résilience et peuvent exécuter l'IA en temps réel sur un grand nombre de transactions, le tout dans le cadre d'un cloud hybride transparent. Le nouveau processeur Power10 d'IBM et la plateforme IBM Power E1080 d'entreprise basée sur Power10 offrent un éventail d'innovations qui répondent à ces exigences de façons inédites.

Le nouveau processeur Power10

La nouvelle architecture et le nouveau processeur Power10 d'IBM présentent d'importantes nouvelles technologies qui aideront les entreprises à gérer leurs charges de travail exigeantes en traitement, mémoire et bande passante, notamment de nouvelles technologies pour l'inférence rapide en IA sur la puce sans matériel supplémentaire, basées sur un moteur MMA (Matrix Math Accelerator) spécialement conçu et intégré.

Du point de vue de la sécurité, Power10 implémente le chiffrement de la mémoire sans dégradation des performances (par opposition au chiffrement de la mémoire basé sur un logiciel), fournit une sécurité des conteneurs co-optimisée au niveau matériel et logiciel pour leur isolement ; il inclut des fonctionnalités de sécurité pour anticiper la capacité imminente de l'informatique quantique à briser les clés de chiffrement traditionnelles.

L'évolutivité avec Power10 est portée à de nouveaux niveaux grâce à plusieurs innovations en bande passante. IBM a amélioré la technologie de connectivité POWER AXON et ajouté l'OMI (Open Memory Interface), tous deux s'exécutant à 32 GT/s. L'interface Power10 AXON permet de connecter jusqu'à 16 sockets pour former un système important et évolutif. L'OMI communique avec la mémoire DRAM DDR4 via 16 ports DDR par socket, fournissant une bande passante pouvant atteindre 409 Go/s par socket. Ces deux interfaces peuvent être utilisées pour fournir des solutions de traitement très flexibles, et même personnalisables.

Il s'agit du premier processeur 7 nanomètres d'IBM qui est trois fois plus efficace que l'IBM Power9 en termes de puissance de traitement (nombre d'utilisateurs, nombre de transactions) et d'énergie.[11] IBM étant axé sur le cloud hybride, cela se traduit directement par un encombrement moindre dans le datacenter et une réduction significative de la consommation énergétique. La puce comporte 15 cœurs de processeur et Power10 prendra en charge PCI Gen5, qui commence à émerger dans le secteur.

L'IBM Power E1080

L'IBM Power E1080 est la première plateforme d'entreprise d'IBM créée avec le processeur Power10. Le système peut s'étendre à 16 processeurs et est particulièrement axé sur les principales considérations informatiques des entreprises qui doivent répondre aux exigences de l'entreprise numérique.

Sécurité

Pour assurer une sécurité constante sans pénalités, IBM a intégré le chiffrement dans le processeur Power10. Les données sont ainsi chiffrées sans compromettre les performances du système. Le système a en outre été équipé de fonctions de sécurité supplémentaires afin de le protéger contre les attaques ROP (Return-Oriented Programming), une technique dans laquelle un pirate peut exécuter un code malveillant en présence de défenses de sécurité. Le Power E1080 assure une

[11] La multiplication par 3 des performances est basée sur une analyse d'ingénierie pré-silicium des environnements Integer, Enterprise et Floating Point sur une offre de serveur à 2 sockets POWER10 avec des modules à 2 x 30 cœurs par rapport à une offre de serveur à 2 sockets POWER9 avec des modules à 2 x 12 cœurs ; les deux modules ont le même niveau d'énergie.

protection avancée des données avec un chiffrement en mémoire transparent, le type de sécurité matérielle pour les données en cours d'utilisation sur lequel repose l'informatique confidentielle ; il comporte quatre fois plus d'accélérateurs de chiffrement cryptographique que son prédécesseur. Les partitions de la plateforme ont amélioré l'isolement et le système est protégé des menaces futures liées à l'informatique quantique via le chiffrement post-quantique et le chiffrement entièrement homomorphe, une technologie dans laquelle les entrées dans le système n'ont pas besoin d'être déchiffrées, ce qui signifie qu'il peut être exécuté par un tiers non fiable sans révéler ces entrées.

Résilience

Pour, IDC la famille de serveurs Power d'entreprise est au niveau AL4 ; en d'autres termes, elle est entièrement tolérante aux pannes et offre donc une disponibilité de 99,999 % ou plus. Avec Power10, l'IBM Power E1080 va plus loin que son prédécesseur, car il offre une fiabilité, une disponibilité et une facilité de maintenance très élevées de la bande passante et de la mémoire, grâce à la nouvelle interface Open Memory. Le processeur peut détecter, isoler et récupérer automatiquement des erreurs logicielles sans indisponibilité ou sans dépendre du système d'exploitation pour gérer les défaillances et autoréparer les erreurs récupérables. Le système présente également des fonctionnalités améliorées de réparation simultanée, telles que des câbles SMP (Sub Miniature Push-on) inter-nœuds pour réduire l'indisponibilité des applications.

Évolutivité et durabilité

En termes d'évolutivité et de durabilité, l'IBM Power E1080 tire énormément parti du fait que la famille de serveurs Power est exceptionnellement bien intégrée, du processeur au firmware, en passant par le système d'exploitation et le matériel, puisque ce sont tous des composants IBM. IBM indique que l'efficacité du logiciel et du conteneur OpenShift de la plateforme est exceptionnelle. Les performances de cette plateforme, dotée du nouveau processeur Power10, sont supérieures de 50 %, avec le même encombrement et la même empreinte énergétique, à celles du Power E980.[12] Cela se traduit également par une consommation énergétique inférieure de 33 % pour la même charge de travail, indique IBM.[13] Cette efficacité accrue permet aux entreprises de réduire considérablement leur empreinte carbone et de consolider potentiellement les charges de travail ; elles réalisent ainsi des gains en matière de coûts matériel et logiciel.

Cloud hybride

Le Power E1080 prend en charge trois environnements d'exploitation, AIX, IBM i et Linux, sur la même plateforme. Il est conçu pour prendre en charge l'adoption du cloud hybride pour ceux-ci. AIX est, bien sûr, le système d'exploitation UNIX entièrement modernisé d'IBM qui continue d'être une plateforme optimale pour la plateforme Power évolutive d'entreprise. IBM i est l'environnement d'exploitation d'IBM qui intègre la base de données et d'autres logiciels d'entreprise au système d'exploitation, simplifiant ainsi grandement la gestion de la plateforme. Pour de nombreuses entreprises de taille moyenne, IBM i constitue le cœur de leurs opérations. AIX et IBM i sont parfaitement adaptés à l'open source, supportent les langages modernes et plébiscités par les développeurs ; ils sont en outre entièrement exploités en tant que cloud hybride. Comme les générations précédentes, le Power E1080 peut également s'exécuter en tout ou partie sur Linux avec les mêmes fonctions de sécurité, de disponibilité et d'évolutivité, ce qui représente une opportunité pour les entreprises de déplacer leurs charges de travail transactionnelles et analytiques vers une plateforme totalement open source.

Les composants logiciels IBM Power ci-après jouent un rôle important, car ils permettent aux entreprises d'exploiter leur plateforme Power via AIX, IBM i et Linux, afin d'assurer une modernisation des charges de travail cloud sécurisée et hautement disponible :

IBM PowerVM

Les charges de travail des serveurs IBM Power sont virtualisées, mobiles et optimisées cloud via PowerVM, qui a récemment bénéficié de plusieurs nouvelles fonctions, notamment la compression et le chiffrement de données LPM (Live Partition Mobility) : quand une partition active fait l'objet d'une migration d'un serveur Power à un autre, ce qui se produit sans indisponibilité, les données sont automatiquement chiffrées et compressées ; il s'agit d'une fonction importante de sécurité et de performance.

IBM PowerVC

PowerVC est l'outil de gestion de la virtualisation qui s'appuie sur OpenStack, qui simplifie la gestion des ressources virtuelles dans les environnements Power. Ce logiciel a récemment bénéficié de nouvelles fonctions, notamment une fonctionnalité d'exportation/importation pour le partage d'images de machines virtuelles entre des datacenters.

[12] Informations fournies par IBM. Basé sur des résultats rPerf publiés pour le cœur Power E980/12 comparés aux mesures rPerf internes IBM (via la même méthodologie) pour le cœur Power E1080/15.

[13] Power9 (12c) : 5081 rPerf @ 16 520 watts (0,31 rPerf/watt), Power10 (15c) : 7998 rPerf @ 17 320 watts (0,46 rPerf/watt) 0,46 / 0,31 = 1,48 rPerf/watt suppl.

IBM PowerSC

PowerSC est la gamme de sécurité de la plateforme qui simplifie la gestion de la sécurité et de la conformité. Il est notamment doté d'une automatisation de la conformité, de la détection d'intrusion de logiciels malveillants et de la gestion de correctifs. Il a bénéficié de plusieurs nouvelles fonctions ou même de nouvelles offres, notamment l'authentification multi-facteur, une autre fonction de sécurité cruciale. Sur IBM Power avec AIX, la sécurité est généralement assurée par une solution complète qui comprend le processeur, le firmware, l'hyperviseur et les innombrables fonctions de sécurité du système d'exploitation pour protéger les données à tous les niveaux.

IBM PowerHA et la solution haute disponibilité et de reprise après incident VM Recovery Manager

PowerHA est une technologie à haute disponibilité qui permet de fournir une disponibilité quasi-continue des applications et d'améliorer la fiabilité du service. Il s'agit d'un contributeur clé d'IBM Enterprise Power. IDC l'a classé comme tolérant aux pannes (AL4) et il a bénéficié de nouvelles fonctions telles que des indicateurs de reprise en ligne améliorés et la vérification inter-clusters (par exemple, pour comparer un cluster de développement à un cluster de test). VM Recovery Manager (VMRM) est une solution simplifiée haute disponibilité et de reprise après incident basée sur la réplication de MV et un redémarrage indépendant du système d'exploitation. Il comprend des agents de surveillance des applications, comme pour DB2, Oracle et SAP HANA.

Cloud Management Console

CMC (Cloud Management Console) offre une vue exhaustive des performances, du stock et de la journalisation de l'infrastructure Power sur et hors site. CMC étant hébergée sur IBM Cloud, les entreprises n'ont plus à gérer de logiciels pour surveiller leur infrastructure. La console simplifie également la gestion des déploiements de cloud hybride, ainsi que la surveillance et la gestion de leur infrastructure.

Enterprise Cloud Edition 2.0

Enterprise Cloud Edition réunit tous les composants clés d'une infrastructure de gestion cloud simplifiée sur PowerVM, y compris PowerSC, MFA, PowerVC, CMC, VMRM, et Aspera. Cette solution permet un déploiement et une gestion rapides d'un cloud privé ; une gestion simplifiée de la sécurité et de la conformité ; une haute disponibilité simplifiée et des transferts accélérés de gros fichiers entre clouds. Enterprise Cloud 2.0 peut être acheté avec AIX 7.2 intégré.

Plateforme d'automatisation Red Hat Ansible

Red Hat Ansible Automation Platform permet une automatisation évolutive et sécurisée de divers aspects des opérations informatiques de l'entreprise, notamment l'allocation de ressources, la gestion du cycle de vie des applications et les opérations réseau. La plateforme est constituée d'Ansible Engine, d'Ansible Tower et d'Ansible Hosted Services. Tous les autres produits de la gamme Red Hat peuvent être intégrés à l'aide de Red Hat Ansible Automation Platform. Red Hat Ansible Automation Platform assure la cohérence dans le datacenter via des méthodes programmatiques pour déployer, gérer et sécuriser les ressources d'infrastructure.

Red Hat OpenShift

Red Hat OpenShift est une plateforme de niveau entreprise, certifiée Kubernetes (une orchestration de conteneur), pour créer, déployer et gérer des applications conteneurisées. Red Hat OpenShift peut être utilisé comme un service intégralement géré sur différents fournisseurs cloud, ou géré par le client via Red Hat OpenShift Container Platform ou Red Hat OpenShift Kubernetes Engine. Il peut être déployé sur site sur des serveurs dédiés, des plateformes de virtualisation (Red Hat Virtualization, VMware, ou Red Hat OpenStack), ou des fournisseurs de cloud majeurs tels qu'IBM Cloud, AWS, Google ou Azure. En outre, Red Hat Advanced Cluster Management for Kubernetes peut être utilisé pour gérer plusieurs clusters et applications Red Hat OpenShift à partir d'une console unique, avec des politiques de sécurité intégrées, les clients pouvant bénéficier d'un cloud hybride ouvert. Red Hat OpenShift est pris en charge sur IBM Power, IBM Z et les plateformes x86 et peut être utilisé avec AIX, IBM i et Linux.

IBM Cloud Paks

Les IBM Cloud Paks sont des produits logiciels de plus en plus populaires, pré-conditionnés dans des conteneurs et hautement intégrés dans divers services OpenShift pour un déploiement rapide et facile sur OpenShift. Les IBM Cloud Paks offrent des outils de développement, des données, des services d'IA et un logiciel middleware open source. Ils s'exécutent sur la plateforme cloud Red Hat OpenShift.

Les Cloud Paks qui sont particulièrement pertinents pour IBM Power sont les suivants :

- ***Cloud Pak for Data :*** aide les clients via des éclairages étendus issus de données et de fonctionnalités IA.
- ***Cloud Pak for Integration :*** se compose d'outils d'intégration pour les données, de services d'application et de services cloud pour l'intégration d'applications, de données, de services cloud et d'API
- ***Cloud Pak for Watson AIOps :*** offre une visibilité, une gouvernance et une automatisation multicloud, compte tenu de l'utilisation courante de déploiements multicloud

Intelligence artificielle

IBM indique que le Power E1080 accélère énormément les performances d'inférence en IA par rapport à son prédécesseur. Aucun matériel spécialisé tel qu'un coprocesseur (GPU, FPGA, ou ASIC) n'est requis. L'inférence a lieu sur un MMA (Matrix Math Accelerator). Chaque cœur de la puce Power10 dispose d'un MMA intégré pour assurer des opérations de calcul matriciel performantes. Ces opérations ont été optimisées sur un large éventail de types de données pour diverses précisions, qui sont importantes pour l'apprentissage profond ; double précision, simple précision et deux types de demi-précision, y compris Bfloat-16 ainsi que Int-16, Int-8 et Int-4 . Les performances en inférence de l'IA ont été injectées dans chaque couche du processeur. Le cache L2 a été quadruplé : les LSU (Load Store Units) et les données SIMD ont doublé. Par conséquent, une charge de travail transactionnelle dotée de composants d'IA intégrés peut exécuter les transactions et l'inférence de l'IA sur le même processeur Power10 sans avoir besoin d'un coprocesseur.

L'inférence sur la puce signifie également que toutes les fonctions de sécurité du processeur et du système protègent les données faisant l'objet de l'inférence. En outre, la plateforme est adaptée à ONNX (Open Neural Network Exchange). ONNX est un écosystème d'IA open source d'entreprises technologiques et d'organismes de recherche qui collaborent pour établir des normes ouvertes pour la représentation des algorithmes et des outils d'IA afin de promouvoir l'innovation et la collaboration dans le secteur de l'IA. Les entreprises équipées d'IBM Power E1080 placent des modèles ONNX sur la plateforme sans les modifier et les exécuter, en tirant parti de ses fonctions RAS pendant l'inférence.

Défis et opportunités

Pour les entreprises

Les plateformes d'entreprise qui exécutent les charges de travail transactionnelles et analytiques clés d'une entreprise ont tendance à être considérées comme des systèmes cloisonnés dans le datacenter, même si elles sont conçues et créées avec des fonctions et des technologies complètes pour l'éviter. Ces plateformes sont souvent « protégées » des nouvelles technologies par une équipe informatique ayant une expertise approfondie du système mais qui hésite à exposer des données, intégrer la plateforme au cloud, exécuter de l'open source sur la plateforme et des modèles d'IA sur des données temps réel. Pour les entreprises, la difficulté consiste à rompre avec cette culture de l'hésitation dès que possible. Il est absolument essentiel qu'elles considèrent les plateformes d'entreprise comme les systèmes ouverts qu'elles sont, ce qui leur permettra de les optimiser pleinement en tant que plateformes de transformation numérique offrant de nouvelles opportunités de revenu. Ces plateformes offrent en outre une opportunité de s'attaquer sérieusement aux problèmes de durabilité, car elles réduisent l'empreinte carbone de l'entreprise. Il devient en outre incontournable de pouvoir exécuter l'IA sur une plateforme d'entreprise sans requérir de coprocesseurs coûteux et énergivores, de plus en plus d'applications clés étant dotées d'une fonctionnalité d'IA.

Pour IBM

Avec la nouvelle plateforme Power E1080, IBM incite une nouvelle fois les entreprises à s'axer sur l'ouverture, le cloud hybride, l'IA et la durabilité dans une plateforme hautement sécurisée, performante et fiable. IBM relève les défis en matière d'innovation en proposant de nouvelles technologies intéressantes qui, dans certains cas, sont révolutionnaires et en avance sur la concurrence : par exemple, le MMA dans le nouveau processeur Power10. L'innovation n'est pas le plus grand défi qu'IBM doit relever. Il s'agit plutôt de faire en sorte que ses clients cessent de considérer leur plateforme d'entreprise non pas comme un système cloisonné, voire un système prudemment ouvert, mais comme une plateforme agressivement intégrée au reste du datacenter ou au cloud. Ils doivent tirer pleinement parti de toutes ses fonctionnalités pour innover et générer plus de revenu avec les données clés y résidant. IBM doit continuer à encourager ses clients à faire preuve d'audace et de créativité avec sa plateforme d'entreprise par le biais de formations, d'incitations et d'études de retour sur investissement.

Conclusion

Les entreprises modernes ont besoin de plateformes informatiques pouvant gérer une volatilité extrême du marché, offrir une sécurité sans faille, s'adapter aisément et de manière durable, fournir la plus grande résilience, exécuter l'IA en temps réel et fonctionner comme un cloud hybride. Le nouveau processeur Power10 d'IBM et la plateforme IBM Power E1080 d'entreprise basée sur celui-ci répondent totalement à ces exigences . Le processeur Power de nouvelle génération d'IBM ne constitue pas une simple avancée : il est tourné vers l'avenir.

Il permet d'utiliser la technologie informatique confidentielle pour assurer un chiffrement matériel qui sécurise les données en cours. La bande passante sur Power10 a été considérablement accrue pour assurer une évolutivité vers 16 sockets. La résilience est encore renforcée via la détection, l'isolement et la récupération automatiques des erreurs logicielles sans indisponibilité ni dépendance vis-à-vis du système d'exploitation. Le MMA sur la puce optimise l'inférence en IA en temps réel sans devoir recourir à un coprocesseur. En outre, entre les solutions Red Hat et le logiciel IBM Cloud, une exploitation intégrale en tant que cloud hybride est bien évidemment acquise. Avec la puce Power10 comme moteur de la nouvelle plateforme Power E1080, IBM s'approche encore plus près d'une informatique d'entreprise idéale, associant ouverture, puissance de traitement, cloud hybride, intelligence artificielle, sécurité, évolutivité, durabilité et fiabilité dans une plateforme unique.

À propos de l'analyste

Peter Rutten

Directeur de recherche, Infrastructure Systems, Platforms and Technologies Group,
Performance Intensive Computing Solutions Global Research Lead, IDC

Peter Rutten est directeur de recherche au sein de la Worldwide Infrastructure Practice d'IDC où il dirige les études sur les plateformes informatiques. M. Rutten est le responsable mondial de la recherche d'IDC sur les solutions informatiques exigeantes en performances et les cas d'utilisation afférents. Cela inclut les études sur l'intelligence artificielle (IA), la modélisation et la simulation (M&S), l'infrastructure BDA (Big Data and Analytics) et les piles de solutions associées. Dans le cadre de l'informatique exigeante en performances, il traite également des systèmes, plateformes et technologies infrastructurels d'informatique de calcul intensif, haut de gamme, accélérée, en mémoire et hétérogène. Il s'occupe en outre des plateformes de traitement dotées de GPU, FPGA, ASIC et autres accélérateurs qui sont déployés dans le cloud comme sur site. Il effectue également des recherches sur les plateformes x86, les mainframes et les systèmes RISC critiques, ainsi que sur leurs environnements d'exploitation (Linux, z/OS, UNIX). M. Rutten étudie également les technologies et plateformes émergentes telles que l'informatique quantique, l'informatique neuromorphique et les technologies potentiellement perturbatrices pour les marchés à infrastructures matures. Dans le cadre de son travail, M. Rutten effectue des analyses quantitatives (estimation de la taille du marché et prévisions le concernant) et qualitatives (basées sur la recherche primaire) ainsi que l'estimation de marché personnalisée pour les clients d'IDC .

En savoir plus sur Peter Rutten

IDC Custom Solutions

Cette publication a été réalisée par IDC Custom Solutions. En tant qu'acteur majeur de la recherche, du conseil et de l'événementiel sur les marchés des technologies de l'information, des télécommunications et des technologies grand public, le groupe Custom Solutions d'IDC aide ses clients à planifier, commercialiser, et vendre leurs produits, et plus généralement à réussir sur les marchés mondiaux. Nous créons des informations exploitables sur les marchés, ainsi que des programmes axés sur du contenu marketing influent produisant des résultats mesurables.

IDC @idc @idc idc.com