Enterprise Strategy Group™
par TechTarget

LIVRE BLANC

Mise en œuvre de solutions optimisées par Ethernet pour l'IA générative

Le rôle essentiel de la gestion de réseau ouverte

Par Bob Laliberte, analyste principal,
Enterprise Strategy Group

Janvier 2024

Ce livre blanc d'Enterprise Strategy Group a été réalisé à la demande de Dell Technologies et est diffusé avec l'autorisation de TechTarget, Inc.

Sommaire

L'infrastructure de l'IA connaît une croissance rapide

À l'échelle mondiale, l'IA générative (GenAI) a déclenché un tsunami d'intérêt et d'activités. Les sites Web TechTarget ont d'ailleurs connu une croissance de plus de 900 % des activités de recherche liées à la GenAI en 2023. Il est important de noter que cette tendance va au-delà du simple intérêt. Les prestataires de services ont été les premiers à adopter cette technologie, beaucoup d'entre eux élargissant leur gamme de services pour inclure des offres de processeur graphique as-a-service. Les grandes entreprises développent quant à elles une infrastructure GenAI privée pour des cas d'utilisation internes tels que l'analytique des consommateurs et la gestion de la chaîne logistique et des stocks. En effet, de nombreux conseils d'administration et cadres dirigeants ont déjà créé des initiatives pour appliquer la GenAI à leurs processus métier. De plus, lors de la dernière conférence Microsoft Ignite, le PDG de Nvidia, Jensen Huang, leader de la GenAI, a prédit que la GenAI aurait un impact significatif, déclarant : : « C'est plus grand que le PC. C'est plus grand que le mobile. Ce sera plus grand qu'Internet. »[1]

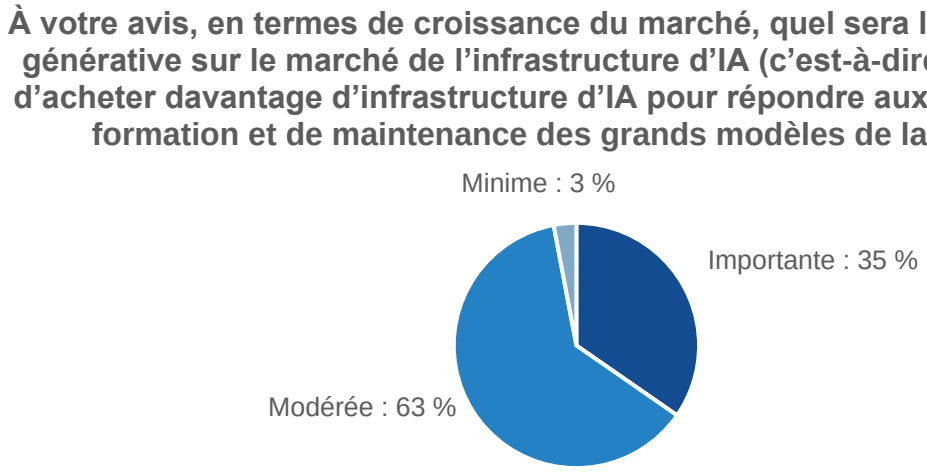
Selon Enterprise Strategy Group (ESG) de TechTarget, il est facile de comprendre pourquoi les organisations sont si impatientes de déployer des solutions de GenAI. D'après l'étude d'ESG, parmi les avantages attendus, l'IA devrait procurer de meilleures informations, accélérer la prise de décision et améliorer le chiffre d'affaires, la rentabilité, l'expérience client et l'efficacité opérationnelle.[2]

Il est également évident que ces initiatives de GenAI exigeront des organisations qu'elles adoptent une nouvelle infrastructure, de nouveaux logiciels et de nouveaux services pour soutenir ces initiatives. Toutefois, ces environnements peuvent considérablement varier, comme l'a souligné Jeff Clarke, vice-président et directeur des opérations chez Dell Technologies. « La GenAI est loin d'être un modèle universel. Il faut une solution de bout en bout, l'infrastructure appropriée, un plan de données, des logiciels et des services qui fonctionnent de manière fluide pour prendre en charge les charges applicatives dans les Clouds, sur site et en périphérie. »

L'étude d'ESG a montré que plus de 9 organisations sur 10 (97 %) estiment que l'infrastructure d'IA connaîtra une croissance importante ou modérée en raison de la GenAI (voir Figure 1).[3] Cette croissance sera nécessaire pour prendre en charge les environnements front-end (utilisateur) et back-end (processeur graphique) afin de garantir des environnements de GenAI robustes.

[1] Source : CRN, « Microsoft Ignite 2023: Nvidia CEO Huang Says Microsoft Is Now 'More Collaborative And Partner-Oriented' », novembre 2023.
[2] Source : Enterprise Strategy Group Complete Survey Results, *Navigating the Evolving AI Infrastructure Landscape*, décembre 2023.
[3] Ibid.

Figure 1. Croissance attendue du marché de l'infrastructure IA avec la GenAI



Source : Enterprise Strategy Group, une division de TechTarget, Inc.

Les organisations ne se contentent pas de faire des recherches sur le sujet : elles prévoient de déployer des environnements de la GenAI, ce qui souligne davantage le souhait d'adopter cette technologie. Les recherches montrent que la grande majorité des personnes interrogées (92 %) prévoient de le faire au cours des 12 prochains mois.[4]

Pour ce faire, les organisations ont besoin d'une infrastructure spécialisée conçue pour gérer les exigences spécifiques de la GenAI, en particulier pour l'environnement de processeur graphique back-end. Toutefois, le déploiement d'une technologie entièrement nouvelle peut présenter des défis à différents niveaux.

Défis liés à l'adoption de nouvelles technologies

Le déploiement d'une nouvelle technologie peut s'avérer difficile pour le département informatique, même s'il s'agit de remplacer simplement une technologie existante. Les nouvelles technologies et/ou architectures peuvent être beaucoup plus difficiles à déployer. Malheureusement, la GenAI nécessite de nouvelles architectures, qui requièrent de nouvelles infrastructures de calcul, de stockage et de réseau, en particulier pour les environnements de processeur graphique back-end. Cela nécessitera non seulement plus d'infrastructure, mais aussi, plus important encore, des systèmes soigneusement conçus pour répondre aux fortes exigences de connectivité entre les clusters de processeurs graphiques. Des connexions Top-of-Rack (ToR) 50 Gigabit Ethernet (GbE) ou 100 GbE classiques avec des liaisons montantes de 400 GbE entraîneraient une congestion et des retards importants pour les grands modèles de langage et mettraient en péril l'ensemble de l'initiative.

Lorsqu'on leur a demandé quels étaient les plus grands défis auxquels les organisations étaient confrontées lors de l'implémentation de solutions d'IA générative, les personnes interrogées ont souligné plusieurs problèmes, notamment l'expertise et les compétences des employés, la complexité technique, l'incapacité à s'intégrer aux systèmes existants, et le coût, parmi de nombreux autres défis liés à la qualité des données, aux considérations éthiques et à la transparence (voir Figure 2).[5]

[4] Ibid.
[5] Source : résultats complets de l'enquête Enterprise Strategy Group, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, août 2023.

Figure 2. Principaux défis liés à la GenAI

Source : Enterprise Strategy Group, une division de TechTarget, Inc.

Il n'est pas surprenant que le principal défi soit le manque de compétences et d'expertise, en particulier pour une technologie émergente telle que l'IA générative. La plupart des organisations ne disposent pas des ressources possédant les compétences requises pour évaluer, concevoir et implémenter une infrastructure de GenAI à grande échelle, en particulier les environnements back-end gourmands en performances.

La complexité technique peut également avoir un impact sur les déploiements de GenAI, car certaines solutions utilisent les technologies propriétaires telles que les réseaux InfiniBand, qui sont généralement réservés aux environnements de calcul haute performance (HPC). Par conséquent, le nombre de ressources possédant les compétences appropriées est limité. Cela s'applique particulièrement aux entreprises et aux hyperscalers qui ont normalisé leurs réseaux Ethernet. Les solutions propriétaires peuvent également être plus difficiles à intégrer dans les plateformes de surveillance ou d'orchestration existantes, ce qui nécessite des compétences, du matériel et des logiciels supplémentaires. Les délais sont un autre facteur à prendre en compte lors de l'utilisation d'une solution propriétaire. Compte tenu des complications survenues ces dernières années concernant la chaîne logistique, les organisations peuvent être réticentes à choisir des solutions disponibles auprès d'un seul fournisseur.

Face à ces défis, les organisations doivent également faire face aux coûts élevés liés à l'implémentation de nouvelles solutions de GenAI, en particulier les solutions propriétaires qui les rendent dépendantes d'un fournisseur spécifique à mesure qu'elles évoluent. Le temps nécessaire à l'évaluation et à la conception d'une solution peut être assez long s'il y a un manque de conceptions et d'architectures de référence.

Les organisations ont besoin d'une infrastructure de GenAI ouverte et robuste

Compte tenu de ces considérations, les organisations doivent se tourner vers des solutions ouvertes pour accélérer le déploiement de l'infrastructure de GenAI. Les organisations devront créer de nouveaux environnements front-end axés sur la facilité d'utilisation et d'accès, qui soutiendront les interactions avec les utilisateurs via une interface Web. L'infrastructure back-end est très différente des environnements traditionnels, voire HPC, et doit prendre en charge de grands modèles de langage (LMM) optimisés par des clusters de processeurs graphiques capables de consommer de grandes quantités de données. Ces environnements d'infrastructure back-end sont essentiels à la réussite d'un projet de GenAI.

Idéalement, ces solutions devraient être :

- **Complètes.** Les organisations qui cherchent à déployer des solutions de GenAI ont besoin de solutions complètes pour les environnements front-end et back-end afin d'accélérer l'adoption. Ces solutions comprennent les ressources de calcul (y compris les clusters de processeurs graphiques), de stockage et de gestion de réseau appropriées pour les deux environnements. Outre l'infrastructure, ces solutions nécessitent des outils complets d'automatisation et de surveillance pour la configuration initiale et la gestion continue, mais également pour faciliter l'optimisation du fabric et le réglage des performances.
- **Hautement performantes.** Pour le réseau, cela signifie déployer des fabrics sans blocage avec une livraison fiable, une bande passante élevée et une faible latence. C'est la raison pour laquelle l'Ultra Ethernet Consortium (UEC) a été créé dans le cadre de la Joint Development Foundation de la Fondation Linux, rassemblant des sociétés pour favoriser une coopération à l'échelle du secteur pour le développement de spécifications Ethernet et d'API logicielles qui permettent aux environnements d'IA d'offrir des performances, une évolutivité, une fiabilité (via le protocole RoCE v2, par exemple) et une interopérabilité de niveau supérieur.[6]
- **Pré-testées et éprouvées.** Pour accélérer l'adoption de ces nouveaux environnements de GenAI, la capacité à déployer une solution complète, testée, à l'efficacité éprouvée, peut aider à éviter les pièges de déploiement courants. L'utilisation de ces solutions élimine une grande partie du temps de recherche, d'analyse et de conception, ce qui permet aux organisations d'atteindre leurs objectifs et de tirer plus rapidement parti de la valeur réelle de leurs environnements de GenAI.
- **Ouvertes et extensibles.** Il s'agit notamment d'employer des composants de série et des fabrics Ethernet plutôt que des technologies de réseau propriétaires. Les environnements de GenAI nécessitent autant de performances réseau que possible, mais avec des normes ouvertes, et non propriétaires. Pour ce faire, l'UEC veillera à ce qu'Ethernet puisse jouer un rôle important dans les environnements de GenAI. En outre, les organisations peuvent tirer parti des systèmes d'exploitation réseau Open Source disponibles dans le commerce, tels que SONiC (Software for Open Networking in the Cloud). Il convient de noter que les projets SONiC et UEC sont hébergés par la Fondation Linux, ce qui facilite la collaboration et l'innovation au sein du secteur.

L'étude d'Enterprise Strategy Group souligne le fait que les organisations qui cherchent à moderniser leurs datacenters sur site ont cité l'utilisation de solutions hyperscale sur site en tant que principale action.[7]

- **Complétées par des services professionnels.** La capacité à accélérer le délai de rentabilisation des solutions de GenAI sera facilitée par des partenaires capables de fournir une expertise et une expérience pertinentes. Cela inclurait la capacité d'effectuer les évaluations appropriées, de mettre au point les conceptions et d'implémenter des solutions en temps opportun. Il peut également s'agir de services entièrement managés, de blueprints techniques ou de conceptions validées.

[6] Ultra Ethernet Consortium.
[7] Source : rapport d'étude Enterprise Strategy Group, *2023 Technology Spending Intentions Survey*, novembre 2022.

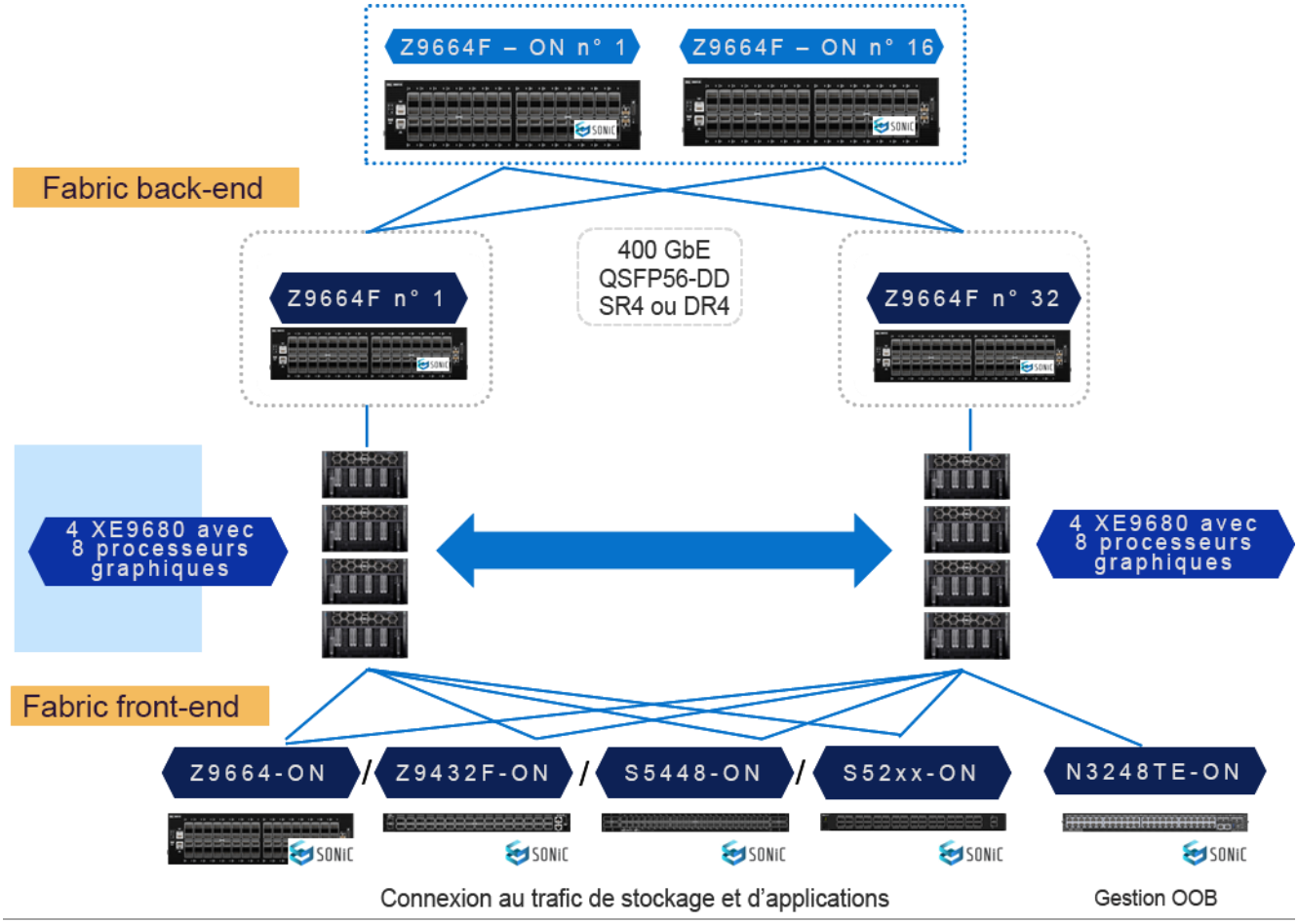
- **Évolutives.** Alors que la plupart des organisations commencent tout juste à utiliser la GenAI, les déploiements initiaux peuvent être limités en taille, mais devront évoluer pour répondre à des exigences accrues. Par conséquent, il sera impératif que l'infrastructure de GenAI et, plus précisément, l'environnement réseau puissent s'étendre pour répondre à ces besoins.
- **À haute efficacité énergétique.** Les solutions basées sur processeurs graphiques nécessitent des quantités considérables d'énergie. C'est pourquoi les organisations doivent prendre toutes les mesures possibles pour réduire la quantité d'énergie consommée. Il convient d'utiliser des puces conçues avec les technologies de dernière génération, qui optimisent les rapports débit/puissance. Les commutateurs haut débit nécessitent moins d'espace rack, d'énergie et de câblage, ce qui en fait une solution plus économique et plus respectueuse de l'environnement. Outre la réduction de la consommation énergétique, la possibilité de fournir des rapports de développement durable aidera également les équipes d'opérations et de gestion.
- **Software-driven.** L'accent mis sur les logiciels accélère le rythme de l'innovation, en particulier pour les logiciels développés dans des environnements ouverts : en effet, ils ne reposent pas sur un seul fournisseur, mais plutôt sur des dizaines d'organisations qui contribuent à leur innovation.

Dell Technologies fournit des solutions ouvertes optimisées par Ethernet pour la GenAI

Dell Technologies fournit des solutions d'infrastructure complètes et ouvertes pour les environnements d'IA, de modélisation et de HPC depuis un certain nombre d'années. Notre société tire parti de son expérience pour mettre en place des solutions d'infrastructure de GenAI pour les environnements front-end (trafic applicatif, accès au stockage, réseau général) et back-end (fabric de processeur graphique) qui incluent le calcul, le stockage et la gestion de réseau.

L'une des clés pour mettre en place une solution de GenAI hautes performances est un fabric réseau d'IA éprouvé et ouvert, comme illustré à la Figure 3.

Figure 3. Solutions complètes de fabric réseau d'IA



Source : Dell Technologies

Les solutions de GenAI de Dell Technologies incluent :

- **Des systèmes de calcul modulaires.** Grâce aux serveurs Dell PowerEdge XE et à l'expérience de la société sur le marché de l'IA, de la modélisation et du HPC, ces serveurs sont optimisés pour l'accélération dans de tels environnements. Dell propose des options de refroidissement par air ou liquide, ainsi que différents nombres de processeurs graphiques, et met un accent particulier sur l'inférence ou l'entraînement des LLM. Notre société dispose ainsi du format et de la solution hautes performances capables de répondre à vos besoins en matière de calcul de GenAI. Les environnements de calcul font partie d'une solution de conception et d'architecture validée pour la GenAI.
- **Un stockage axé sur l'IA.** Dell propose différentes options de stockage en fonction des exigences des charges applicatives, notamment les solutions PowerScale, Elastic Cloud Storage et ObjectScale. Le stockage PowerScale OneFS basé sur Ethernet permet aux lectures et écritures en streaming d'accéder rapidement aux données des charges applicatives d'IA et améliore la capacité de modélisation de l'IA. Dell précise que PowerScale a été testé sur le terrain auprès de plus de 1 000 clients exécutant des charges applicatives sur processeur graphique. C'est pourquoi il existe de nombreuses solutions de conception validée Dell Validated Design, basées sur ces expériences. La vaste gamme d'options est également certifiée Energy Star.

- **Des fabrics Ethernet de nouvelle génération.** Ce matériel de réseau ouvert, axé sur le commutateur Dell PowerSwitch et doté d'une puce de nouvelle génération, telle que Broadcom Tomahawk 4, peut fournir jusqu'à 51,2 Tbit/s avec une mise en mémoire tampon partagée des paquets. Disponibles dans le commerce sous le nom PowerSwitch série Z, les commutateurs 64 ports Z9664F-ON et 32 ports Z9432F-ON peuvent évoluer pour prendre en charge des milliers de nœuds. En outre, Dell Technologies est membre de l'UEC et contribuera à l'extension de l'applicabilité d'Ethernet aux environnements de GenAI.
- **Des architectures software-driven.** Dell Technologies s'engage toujours à fournir des solutions de gestion de réseau ouverte pour les systèmes d'exploitation réseau, l'orchestration et la surveillance dans les environnements de GenAI. Pour le système d'exploitation réseau, Dell Technologies a adopté et renforcé SONiC, en fournissant le support mondial, l'évolutivité et les fonctionnalités requis par les grandes entreprises. La dernière version d'Enterprise SONiC Distribution by Dell Technologies (version 4.2) offre une prise en charge avancée des environnements d'IA qui inclut RDMA over Converged Ethernet version 2 (RoCE v2), le hachage amélioré et la commutation directe. La prochaine version 4.3 apporte des améliorations pour l'équilibrage de charge et le mappage. Toutes les versions de SONiC sont testées et validées sur l'ensemble de la gamme série Z. Elles sont également testées par rapport à l'écosystème de partenaires d'applications tierces de Dell.
- **Des services pour accélérer l'adoption et l'optimisation.** En plus d'un support mondial 24x7, Dell Technologies peut compter sur des experts en services professionnels disposant d'une expérience éprouvée pour permettre aux organisations d'évaluer, de concevoir et d'implémenter correctement des solutions complètes de GenAI. Leur capacité à comprendre non seulement le réseau, mais aussi les domaines de calcul et de stockage, accélère le processus de conception et réduit le risque de problèmes de compatibilité. Ces conceptions validées couvrent à la fois l'inférence et la personnalisation des modèles, et il existe des services pour la préparation et l'ingestion des données pour les opportunités de GenAI. Dell propose également des services managés pour exploiter ces environnements d'IA.
- **Une concentration sur le développement durable.** Le déploiement d'environnements de GenAI à grande échelle nécessite d'importantes ressources d'alimentation. Les commutateurs haut débit Dell en mode dérivation nécessitent moins d'espace rack, d'alimentation et de câblage. L'utilisation des puces de toute dernière technologie permet de bénéficier de la meilleure efficacité énergétique possible pour les serveurs, la gestion de réseau et les solutions de stockage. En se concentrant sur l'efficacité énergétique, les organisations peuvent réduire leurs coûts et leur consommation électrique.

Grâce à ces intégrations, Dell Technologies occupe une position idéale pour fournir des solutions complètes d'infrastructure de GenAI pour les environnements back-end et front-end.

Conclusion

La hausse de l'intérêt et des activités liés à la GenAI pousse les organisations à évaluer des solutions pour leurs propres environnements. Cependant, sa popularité étant récente, la plupart des équipes informatiques n'ont pas l'expertise ou l'expérience nécessaires pour implémenter une solution en temps opportun. En outre, il faut avouer que ces infrastructures de GenAI, qui nécessitent des architectures et des technologies nouvelles, sont très complexes. Elles doivent être soigneusement conçues et fournir un système équilibré, de sorte qu'il peut être très risqué de se procurer des composants disparates et d'essayer de les assembler. Les organisations doivent donc établir des partenariats stratégiques pour acquérir les compétences et les solutions étroitement intégrées nécessaires à la réussite d'un environnement de GenAI.

Toutefois, les organisations doivent se méfier des solutions complètes qui les enferment dans des technologies propriétaires, en particulier lorsque ces environnements évoluent. Les solutions ouvertes peuvent fournir innovation, flexibilité et rentabilité pour les environnements de GenAI à grande échelle. Toutefois, pour garantir la robustesse d'un environnement, il est également essentiel de s'assurer que ces solutions ouvertes sont entièrement testées, validées et prises en charge.

Dell Technologies fournit des solutions complètes de GenAI qui intègrent l'ensemble de l'infrastructure et des logiciels, y compris l'orchestration et la gestion pour les environnements front-end et back-end. Elles intègrent également des systèmes ouverts de calcul, de stockage et de gestion de réseau. De plus, les organisations peuvent utiliser des services managés, des services professionnels, ainsi que des conceptions et des architectures entièrement validées qui incluent l'écosystème de partenaires Dell. Ces solutions à la fois complètes et modulaires permettent aux organisations d'accélérer le déploiement et la valeur des solutions de GenAI, tout en réduisant les risques et en garantissant une meilleure efficacité opérationnelle.

À propos d'Enterprise Strategy Group
Enterprise Strategy Group de TechTarget fournit des informations ciblées et exploitables sur le marché, des recherches sur la demande, des services consultatifs d'analystes, des conseils en matière de stratégie GTM, des validations de solutions et du contenu personnalisé pour soutenir l'achat et la vente de technologies d'entreprise.

contact@esg-global.com
www.esg-global.com