Enterprise Strategy Group™
di TechTarget

WHITE PAPER

Attuazione di soluzioni basate su Ethernet per l'AI generativa

L'importanza delle reti aperte

Di Bob Laliberte, Principal Analyst, Enterprise Strategy Group

Gennaio 2024

Questo white paper di Enterprise Strategy Group è stato commissionato da Dell Technologies ed è distribuito su licenza di TechTarget, Inc.

Sommario

La rapida espansione dell'infrastruttura AI

A livello globale, l'AI generativa (GenAI) ha scatenato un'ondata di interesse e attività. Non a caso, nel 2023 i siti web di TechTarget hanno registrato una crescita di oltre il 900% delle attività di ricerca relative alla GenAI. È importante sottolineare che non si tratta di semplice interesse. I fornitori sono stati tra i primi ad adottare questa tecnologia: molti hanno ampliato il proprio portafoglio di servizi per includere offerte GPU as-a-Service, mentre le grandi imprese stanno ampliando l'infrastruttura GenAI privata per casi d'uso interni, come l'analisi dei consumatori e la gestione di supply chain e inventario. Numerosi consigli di amministrazione e dirigenti aziendali hanno già avviato iniziative per applicare la GenAI ai processi di business. Inoltre, in occasione dell'ultima conferenza Microsoft Ignite, Jensen Huang, Amministratore delegato di Nvidia e responsabile della GenAI, ha previsto che questa tecnologia avrà un impatto significativo. Ha affermato: "È una rivoluzione più importante del PC. Più importante delle tecnologie mobili. Sarà più importante di Internet".[1]

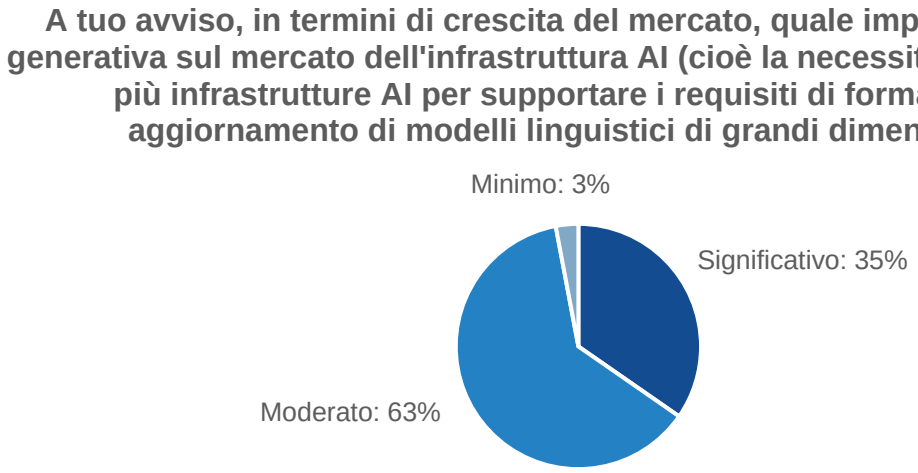
Secondo l'Enterprise Strategy Group (ESG) di TechTarget, è facile capire perché le organizzazioni siano così interessate a implementare soluzioni GenAI. Secondo la ricerca di ESG, tra i vantaggi attesi dall'AI rientrano informazioni approfondite di migliore qualità, incremento di entrate e redditività, processi decisionali più rapidi, esperienze cliente ed efficienza operativa migliorate.[2]

Inoltre, è chiaro che per supportare tali iniziative, le organizzazioni dovranno adottare nuovi software, infrastrutture e servizi. Tuttavia, questi ambienti possono variare notevolmente, come ha osservato Jeff Clarke, Vice Chairman e Chief Operating Officer presso Dell Technologies. "La GenAI è ben lungi dal riuscire a soddisfare tutti i tipi di esigenze. Richiede una soluzione end-to-end, l'infrastruttura adeguata, un piano dati, software e servizi che funzionino perfettamente per supportare i carichi di lavoro su più cloud, on-premise e nell'edge."

La ricerca di ESG ha dimostrato che oltre 9 organizzazioni su 10 (97%) ritengono che l'AI generativa favorirà una crescita significativa o moderata dell'infrastruttura AI (vedere Figura 1).[3] Un passaggio necessario per supportare ambienti front-end (utente) e back-end (GPU) allo scopo di garantire ambienti GenAI affidabili.

[1] Fonte: CRN, "Microsoft Ignite 2023: Nvidia CEO Huang Says Microsoft Is Now 'More Collaborative And Partner-Oriented'", novembre 2023.
[2] Fonte: Enterprise Strategy Group, risultati completi del sondaggio, *Navigating the Evolving AI Infrastructure Landscape*, dicembre 2023.
[3] Ibid.

Figura 1. Previsione di crescita nel mercato dell'infrastruttura AI derivante da GenAI



Fonte: Enterprise Strategy Group, una divisione di TechTarget, Inc.

Ribadendo ulteriormente la volontà di adottare l'AI generativa, le organizzazioni non si fermano alla semplice raccolta di informazioni sull'argomento, ma pianificano l'implementazione di ambienti GenAI: secondo la ricerca, la stragrande maggioranza degli intervistati (92%) prevede di farlo nei 12 mesi successivi.[4]

Per raggiungere questo obiettivo, le organizzazioni hanno bisogno di un'infrastruttura specializzata progettata per gestire i requisiti specifici della GenAI, in particolare per l'ambiente GPU di back-end. Tuttavia, l'implementazione di una tecnologia completamente nuova può presentare sfide a diversi livelli.

Le sfide del passaggio a una nuova tecnologia

L'implementazione di una nuova tecnologia può costituire una sfida per l'IT, anche se si tratta semplicemente di sostituire quella esistente. Le tecnologie e/o architetture di ultima generazione possono essere molto più difficili da implementare. Purtroppo, l'AI generativa richiede nuove architetture, che a loro volta richiedono nuove infrastrutture di elaborazione, storage e rete, soprattutto per gli ambienti GPU di back-end. Ciò significa che oltre a un'infrastruttura più estesa, saranno richiesti anche e soprattutto sistemi progettati con attenzione per soddisfare gli enormi requisiti in termini connettività dei cluster di GPU. Le tipiche connessioni da 50 Gigabit Ethernet (GbE) o 100 GbE top-of-rack (ToR) con uplink da 400 GbE causerebbero congestioni e ritardi significativi per i modelli linguistici di grandi dimensioni e metterebbero a rischio l'intera iniziativa.

Alla domanda sulle maggiori sfide che le organizzazioni devono affrontare durante l'implementazione di soluzioni GenAI, gli intervistati hanno evidenziato diversi problemi, tra cui l'esperienza e le competenze dei dipendenti, la complessità tecnica, l'impossibilità di realizzare l'integrazione con i sistemi esistenti o legacy, nonché i costi, oltre a molte altre difficoltà legate a qualità dei dati, considerazioni etiche e trasparenza (vedere Figura 2).[5]

[4] Ibid.
[5] Fonte: risultati completi del sondaggio di Enterprise Strategy Group, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, agosto 2023.

Figura 2. Le principali sfide dell'AI generativa

Fonte: Enterprise Strategy Group, una divisione di TechTarget, Inc.

Non dovrebbe sorprendere che la sfida principale sia la mancanza di competenze ed esperienza, soprattutto nel caso di una tecnologia emergente come l'AI generativa: la maggior parte delle organizzazioni non dispone delle risorse con le conoscenze necessarie per valutare, progettare e implementare un'infrastruttura GenAI su larga scala, in particolare gli ambienti di back-end ad alta intensità di prestazioni.

La complessità tecnica può influire anche sulle implementazioni GenAI, poiché alcune soluzioni sfruttano tecnologie proprietarie, come le reti InfiniBand, generalmente riservate agli ambienti HPC (High Performance Computing). Di conseguenza, le risorse con le competenze adeguate sono limitate e questo vale in particolare per aziende e hyperscaler che di norma utilizzano reti Ethernet. Inoltre, l'integrazione delle soluzioni proprietarie nelle piattaforme di orchestration o monitoraggio esistenti può risultare più difficile poiché richiede competenze, hardware e software aggiuntivi. Un altro elemento da prendere in considerazione quando si utilizza una soluzione proprietaria sono i lead time: in ragione delle complicazioni degli ultimi anni nella supply chain, le organizzazioni potrebbero esitare a scegliere le soluzioni offerte da un unico fornitore.

A causa di queste sfide, inoltre, le aziende devono anche affrontare i costi elevati associati all'implementazione di nuove soluzioni GenAI, in particolare quelle proprietarie che le vincolano a un vendor specifico man mano che vengono scalate. Il tempo necessario per valutare e progettare una soluzione può essere piuttosto lungo in mancanza di progetti e architetture di riferimento.

Le organizzazioni richiedono un'infrastruttura GenAI aperta e robusta

Alla luce di queste considerazioni, le organizzazioni devono cercare soluzioni aperte per accelerare il deployment dell'infrastruttura GenAI. Dovranno creare nuovi ambienti front-end che consentano le interazioni degli utenti tramite un'interfaccia web-based e che siano incentrati sulla facilità d'uso e di accesso. L'infrastruttura di back-end è molto diversa dagli ambienti tradizionali o da quelli HPC e dovrebbe supportare modelli linguistici di grandi dimensioni (LMM) basati su cluster GPU in grado di ingerire grandi quantità di dati. Questi ambienti con l'infrastruttura di back-end sono fondamentali per il successo di un progetto GenAI.

Idealmente, queste soluzioni dovrebbero essere:

- **Complete.** Le organizzazioni che intendono implementare soluzioni GenAI hanno bisogno di un'offerta completa per ambienti front-end e back-end al fine di accelerarne l'adozione. Questa deve includere le adeguate risorse di elaborazione (inclusi i cluster GPU), storage e di rete per entrambi gli ambienti. Oltre all'infrastruttura, queste soluzioni hanno bisogno di strumenti completi di automazione e monitoraggio non solo per la configurazione iniziale e la gestione continua, ma che siano anche in grado di supportare l'ottimizzazione della fabric e la messa a punto delle prestazioni.
- **A prestazioni elevate.** A livello di rete, ciò significa implementare fabric non bloccanti con consegna affidabile, larghezza di banda elevata e bassa latenza. Questo è il motivo che ha portato alla creazione del Consorzio Ultra Ethernet (UEC) come parte della Linux Foundation's Joint Development Foundation, che riunisce le aziende allo scopo di stabilire la cooperazione a livello di settore per lo sviluppo di specifiche Ethernet e API software, in grado di potenziare gli ambienti AI con prestazioni, scalabilità, affidabilità (ad esempio, con il protocollo RoCE v2) e interoperabilità di livello superiore.[6]
- **Pre-testate e comprovate.** Per accelerare l'adozione di questi nuovi ambienti GenAI, la possibilità di implementare una soluzione completa testata e di comprovata efficacia può contribuire a evitare le difficoltà più comuni del deployment. Queste soluzioni eliminano gran parte del tempo necessario per ricerca, analisi e progettazione, permettendo alle organizzazioni di raggiungere più rapidamente gli obiettivi e il valore reale derivanti dagli ambienti GenAI.
- **Aperte ed estensibili.** Grazie all'utilizzo di silicio commerciale e fabric Ethernet invece di tecnologie di rete proprietarie. Gli ambienti GenAI richiedono il massimo delle prestazioni di rete possibili, ma da standard aperti, non proprietari. A tal fine, l'UEC garantirà che Ethernet svolga un ruolo significativo negli ambienti GenAI. Inoltre, le organizzazioni possono sfruttare i sistemi operativi di rete open source disponibili in commercio, come SONiC (Software for Open Networking in the Cloud). È necessario sottolineare che sia SONiC sia UEC sono ospitati dalla Linux Foundation, in modo da facilitare la collaborazione e l'innovazione nel settore.

La ricerca di Enterprise Strategy Group evidenzia come le organizzazioni che desiderano modernizzare i data center on-premise abbiano dichiarato che l'adozione di soluzioni iperscalabili on-premise ha la massima priorità.[7]

- **Potenziate con servizi professionali.** I partner in grado di fornire competenze ed esperienza pertinenti favoriranno l'accelerazione del time-to-value per le soluzioni GenAI; tali competenze includono la capacità di effettuare le appropriate valutazioni, di concepire i progetti e implementare soluzioni in modo tempestivo. Inoltre potrebbero includere anche servizi completamente gestiti, blueprint tecnici o progettazioni convalidate.

[6] Consorzio Ultra Ethernet.
[7] Fonte: report di ricerca Enterprise Strategy Group, *2023 Technology Spending Intentions Survey*, novembre 2022.

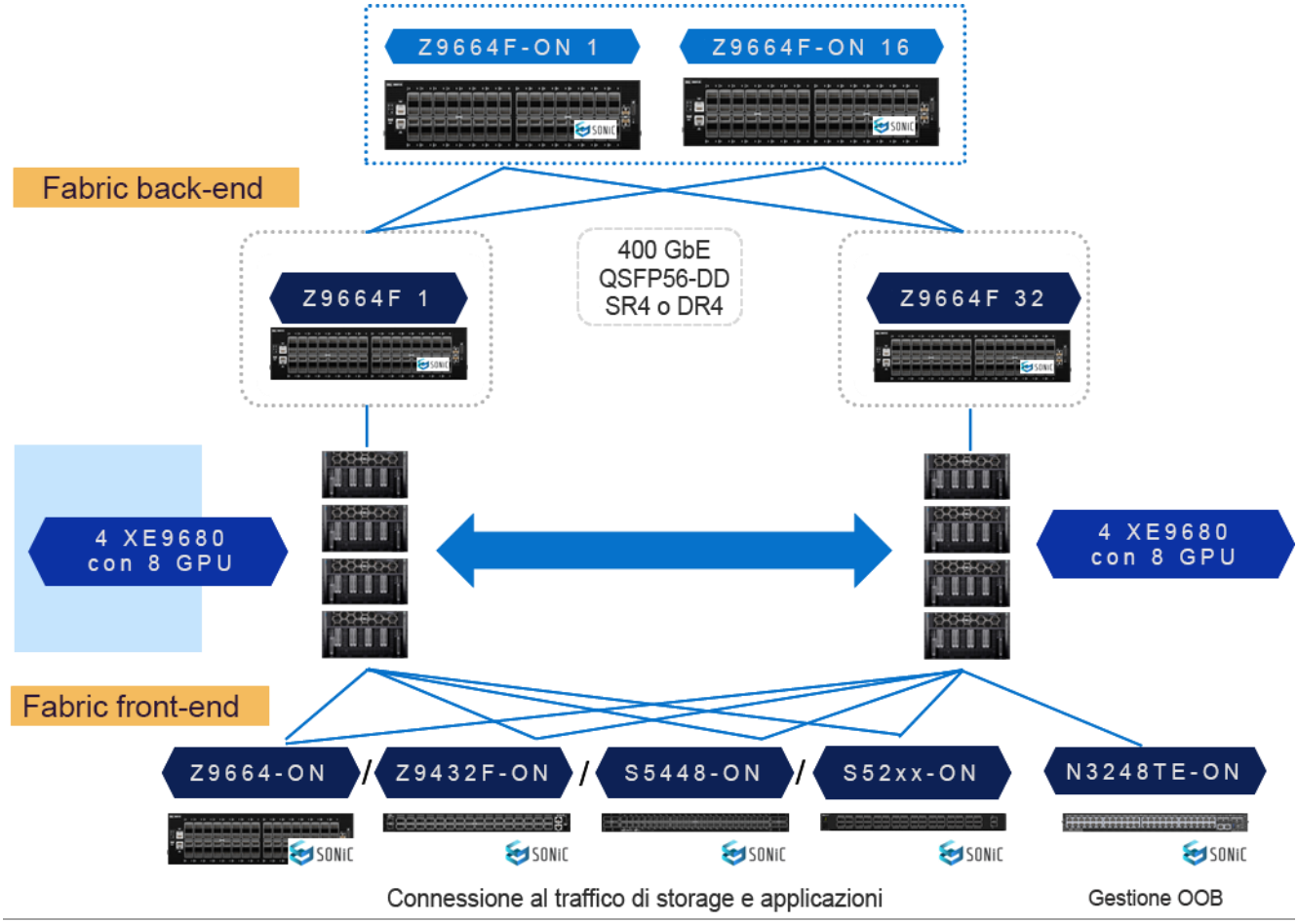
- **Scalabili.** Poiché la maggioranza delle organizzazioni ha appena iniziato a utilizzare l'AI generativa, le prime implementazioni potrebbero essere di dimensioni limitate, ma dovranno essere scalabili per soddisfare requisiti sempre crescenti. Pertanto, sarà fondamentale che l'infrastruttura GenAI e, più specificamente, l'ambiente di rete possano espandersi per soddisfare queste esigenze.
- **Efficienti dal punto di vista energetico.** Le soluzioni basate su GPU richiedono enormi quantità di energia, di conseguenza le organizzazioni devono adottare tutte le misure possibili per ridurre il consumo energetico. È necessario utilizzare la tecnologia al silicio di ultima generazione che ottimizza il rapporto produttività/potenza. Gli switch a velocità più elevata occupano meno spazio sul rack, utilizzano meno energia e cablaggio, offrendo una soluzione più economica ed ecologica. Oltre alla riduzione del consumo energetico, la possibilità di produrre report di sostenibilità sarà di aiuto anche ai team operativi e di gestione.
- **Basate su software.** Concentrarsi sul software accelera il ritmo dell'innovazione, soprattutto se è sviluppato in ambienti aperti, in quanto non si basa su un unico fornitore, ma potenzialmente su decine di organizzazioni che contribuiscono alla sua modernizzazione.

Dell Technologies offre soluzioni GenAI aperte basate su Ethernet

Da diversi anni, Dell Technologies offre soluzioni per l'infrastruttura complete e aperte per ambienti AI, di modellazione e HPC. Sfrutta la sua consolidata esperienza per fornire soluzioni per l'infrastruttura GenAI sia per ambienti front-end (traffico delle applicazioni, accesso allo storage, rete generale) sia back-end (fabric GPU), che includono elaborazione, storage e networking.

Uno degli elementi chiave per abilitare una soluzione GenAI a prestazioni elevate è una fabric di rete AI comprovata e aperta, come illustrato nella Figura 3.

Figura 3. Soluzioni complete per fabric di rete AI



Fonte: Dell Technologies.

Le soluzioni GenAI Dell Technologies includono:

- **Sistemi di elaborazione modulari.** Basati sui server Dell PowerEdge XE e sull'esperienza dell'azienda al servizio del mercato di AI, modellazione e HPC, questi server sono ottimizzati in termini di accelerazione per tali ambienti. Con le opzioni per il raffreddamento ad aria o a liquido, nonché il numero di GPU, unite a un'attenzione particolare all'inferenza o al training degli LLM, Dell offre la soluzione con il fattore di forma giusto e a prestazioni elevate per soddisfare le tue esigenze di elaborazione GenAI. Gli ambienti di elaborazione fanno parte di una soluzione di progettazione e architettura convalidata per l'AI generativa.
- **Storage incentrato sull'AI.** Dell offre una gamma di opzioni di storage disponibili in base ai requisiti dei carichi di lavoro, tra cui soluzioni PowerScale, Elastic Cloud Storage e ObjectScale. Lo storage PowerScale OneFS basato su Ethernet consente letture e scritture in streaming per accedere rapidamente ai dati dei carichi di lavoro AI e migliora la capacità di modellazione AI. Dell dichiara che PowerScale è stato testato sul campo con oltre 1.000 clienti che lo hanno usato per eseguire carichi di lavoro GPU. Di conseguenza, esistono numerose progettazioni convalidate Dell basate su tali esperienze. Le numerose opzioni disponibili hanno ricevuto la certificazione Energy Star.

- **Fabric Ethernet di nuova generazione.** Incentrato su Dell PowerSwitch e dotato silicio di nuova generazione, come Tomahawk 4 di Broadcom, questo hardware di rete aperta può fornire fino a 51,2 Tbps con buffering dei pacchetti condiviso. Disponibili in commercio come PowerSwitch serie Z, gli switch Z9664F-ON a 64 porte e Z9432F-ON a 32 porte possono essere scalati per supportare migliaia di nodi. Inoltre, Dell Technologies fa parte dell'UEC e contribuirà ad ampliare l'applicabilità di Ethernet per alimentare gli ambienti GenAI.
- **Architetture basate su software.** Dell Technologies ribadisce il proprio impegno a fornire soluzioni di rete aperte per sistemi operativi di rete, orchestration e monitoraggio in ambienti GenAI. Per il sistema operativo di rete, Dell Technologies ha adottato e rafforzato SONiC, offrendo supporto globale, scalabilità e le funzioni richieste dalle grandi imprese. L'ultimo Enterprise SONiC Distribution by Dell Technologies (versione 4.2) offre supporto avanzato per ambienti AI con RDMA su Converged Ethernet versione 2 (RoCE v2), hash avanzato e switch cut-through. La prossima versione 4.3 offrirà miglioramenti per il bilanciamento e il mapping del carico. Tutte le versioni di SONiC sono testate e convalidate per tutto il portafoglio della serie Z. Le versioni vengono testate anche nell'ecosistema dei partner di applicazioni di terze parti Dell.
- **Fornitura di servizi per accelerare adozione e ottimizzazione.** Oltre al supporto globale 24 ore su 24, 7 giorni su 7, Dell Technologies dispone di esperti di servizi professionali con esperienza comprovata per consentire alle organizzazioni di valutare, progettare e implementare correttamente soluzioni GenAI complete. La loro capacità di comprendere non solo la rete, ma anche gli ambiti di elaborazione e storage accelera il processo di progettazione e riduce il rischio che emergano problemi di compatibilità. Queste progettazioni convalidate coprono sia l'inferenza che la personalizzazione dei modelli e mettono a disposizione servizi per la preparazione e l'acquisizione dei dati per le pipeline GenAI. Dell offre anche servizi gestiti per il funzionamento degli ambienti AI.
- **Attenzione alla sostenibilità.** L'implementazione di ambienti GenAI su larga scala richiede considerevoli risorse energetiche. Gli switch Dell con velocità più elevata in modalità breakout richiedono meno spazio su rack, alimentazione e cablaggio. Grazie alla più recente tecnologia del silicio, server, reti e soluzioni di storage non sono mai stati così efficienti dal punto di vista energetico. Fare attenzione all'efficienza energetica consente alle organizzazioni di ridurre i costi e il consumo energetico.

Grazie a questi complementi, Dell Technologies è in grado di fornire soluzioni complete per l'infrastruttura GenAI per ambienti back-end e front-end.

Conclusioni

L'aumento dell'interesse verso l'AI generativa e le attività intraprese in questo ambito spingono le organizzazioni a valutare soluzioni per i propri ambienti. Tuttavia, a causa della sua recente popolarità, la maggior parte dei team IT non possiede le competenze o l'esperienza necessarie per implementare una soluzione in modo tempestivo. Inoltre, le infrastrutture GenAI, che richiedono nuove architetture e tecnologie, presentano un'elevata complessità: devono essere progettate con cura e fornire un sistema bilanciato, quindi cercare di procurarsi i singoli componenti e poi combinarli può essere molto rischioso. Per questo motivo, le organizzazioni devono collaborare in modo strategico per acquisire competenze e soluzioni altamente integrate al fine di garantire il corretto funzionamento dell'ambiente GenAI.

Tuttavia, devono fare attenzione alle soluzioni complete che le vincolano a una tecnologia proprietaria, soprattutto considerando la scalabilità di questi ambienti, mentre le soluzioni aperte offrirono innovazione, flessibilità e convenienza per gli ambienti GenAI su larga scala. In ogni caso, per assicurare la stabilità degli ambienti, è fondamentale anche garantire che le soluzioni aperte siano completamente testate, convalidate e supportate.

Dell Technologies fornisce soluzioni GenAI complete che comprendono l'intera infrastruttura e il software, incluse l'orchestration e la gestione per ambienti front-end e back-end, nonché elaborazione, storage e reti aperte. Inoltre, le organizzazioni possono avvalersi di servizi gestiti, Professional Services e progettazioni e architetture completamente convalidate che includono l'ecosistema dei partner Dell. Queste soluzioni complete ma modulari consentono alle organizzazioni di accelerare il deployment e il valore delle soluzioni GenAI, riducendo al contempo i rischi e garantendo una maggiore efficienza operativa.

About Enterprise Strategy Group
Enterprise Strategy Group di TechTarget fornisce intelligence di mercato mirata e fruibile, ricerche per la domanda, servizi di consulenza da parte di analisti, indicazioni sulla strategia GTM, convalide di soluzioni e contenuto del cliente a supporto dell'acquisto e della vendita di tecnologia aziendale.

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