Enterprise Strategy Group by TechTarget

백서

GenAI를 위한 이더넷 기반 솔루션 지원

오픈 네트워킹의 중요성

작성자: Bob Laliberte, Principal Analyst
Enterprise Strategy Group

2024년 1월

목차

빠르게 성장하고 있는 AI 인프라스트럭처

전 세계적으로 GenAI(Generative AI)가 엄청난 관심을 받으면서 이에 관련된 활동도 크게 증가했다. 실제로 TechTarget 웹사이트에서 관찰한 결과, 2023년에 GenAI와 관련된 검색 활동이 900% 이상 늘어났다. 여기서 중요한 점은 이러한 현상이 단순한 관심 그 이상으로 확대되고 있다는 것이다. 서비스 공급업체는 이 기술의 얼리 어답터로, 많은 업체가 GPU as-a-Service 오퍼링을 포함하도록 서비스 포트폴리오를 확장하고 있으며, 대규모 기업은 소비자 분석, 공급망 및 인벤토리 관리와 같은 내부 활용 사례를 위해 프라이빗 GenAI 인프라스트럭처를 구축하고 있다. 실제로 많은 기업 이사회와 최고위 경영진이 이미 GenAI를 비즈니스 프로세스에 적용하기 위한 이니셔티브를 마련했다. 가장 최근 Microsoft Ignite 컨퍼런스에서 GenAI 분야의 선두 주자인 Nvidia의 CEO Jensen Huang은 GenAI가 중대한 영향을 미칠 것이라 예측하며 이렇게 말했다. "GenAI는 PC보다 더 큽니다. 모바일보다 더 큽니다. 그리고 인터넷보다 더 커질 것입니다."[1]

TechTarget의 ESG(Enterprise Strategy Group)에 따르면 조직이 GenAI 솔루션을 구축하고자 하는 이유는 쉽게 이해할 수 있다. ESG의 연구에 따르면 AI를 통해 얻을 수 있는 이점으로는 통찰력 향상, 매출과 수익성 개선, 더 빠른 의사 결정 속도, 고객 경험 향상, 운영 효율성 향상 등이 있다.[2]

이러한 GenAI 이니셔티브를 지원하려면 조직이 새로운 인프라스트럭처, 소프트웨어 및 서비스를 채택해야 한다는 사실도 명확하다. 그러나 Dell Technologies의 부회장 겸 최고 운영 책임자인 Jeff Clarke는 이러한 환경이 매우 다양할 수 있다고 말한다. "GenAI는 획일적인 모델과는 거리가 멉니다. 이를 위해서는 클라우드, 온프레미스, 엣지 환경 모두에서 워크로드를 지원하기 위해 원활하게 작동하는 포괄적인 솔루션, 적절한 인프라스트럭처, 데이터 요금제, 소프트웨어 및 서비스가 필요합니다."

ESG의 연구에 따르면 조직 10곳 중 9곳 이상(97%)이 GenAI로 인해 AI 인프라스트럭처가 상당한 수준 또는 보통 수준의 성장을 보일 것으로 생각하고 있다(그림 1 참조).[3] 이는 견고한 GenAI 환경을 위해 프런트엔드(사용자) 환경과 백엔드(GPU) 환경을 모두 지원하는 데 필요하다.

---

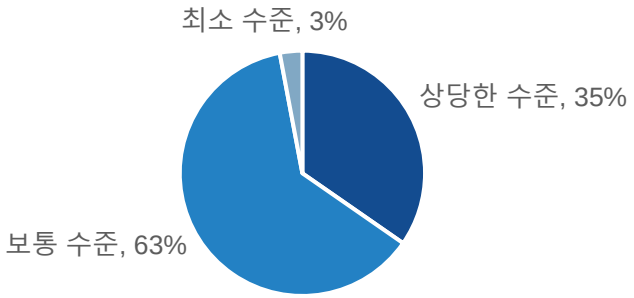
[1] 출처: CRN, "Microsoft Ignite 2023: Nvidia CEO Huang Says Microsoft Is Now 'More Collaborative And Partner-Oriented'", 2023년 11월.

[2] 출처: Enterprise Strategy Group Complete Survey Results, *Navigating the Evolving AI Infrastructure Landscape*, 2023년 12월.

[3] 같은 문서 참조.

그림 1. GenAI로 인한 AI 인프라스트럭처 시장의 성장 예상

대규모 언어 모델의 학습 및 유지 관리 요구 사항을 지원하기 위해 더 많은 AI 인프라스트럭처를 구매해야 하는 경우가 있습니다. 이렇게 시장 성장 측면에서 Generative AI 가 AI 인프라스트럭처 시장에 미칠 영향이 얼마나 클 것이라고



출처: Enterprise Strategy Group, TechTarget, Inc. 산하 사업부

GenAI를 수용하려는 움직임이 더욱 강화됨에 따라 조직들은 단순히 이 분야를 조사하는 것을 넘어 GenAI 환경을 구축할 계획을 세우고 있다. 연구 결과에 따르면 응답자의 대다수(92%)가 향후 12개월 이내에 구축할 계획인 것으로 나타났다.[4]

이를 위해 조직은 GenAI의 특정 요구 사항을 처리하도록 설계된 전문 인프라스트럭처가 필요하며, 특히 백엔드 GPU 환경이 필요하다. 그러나 완전히 새로운 기술을 구축하는 것은 다양한 수준에서 문제를 야기할 수 있다.

새로운 기술로 전환할 때의 당면 과제

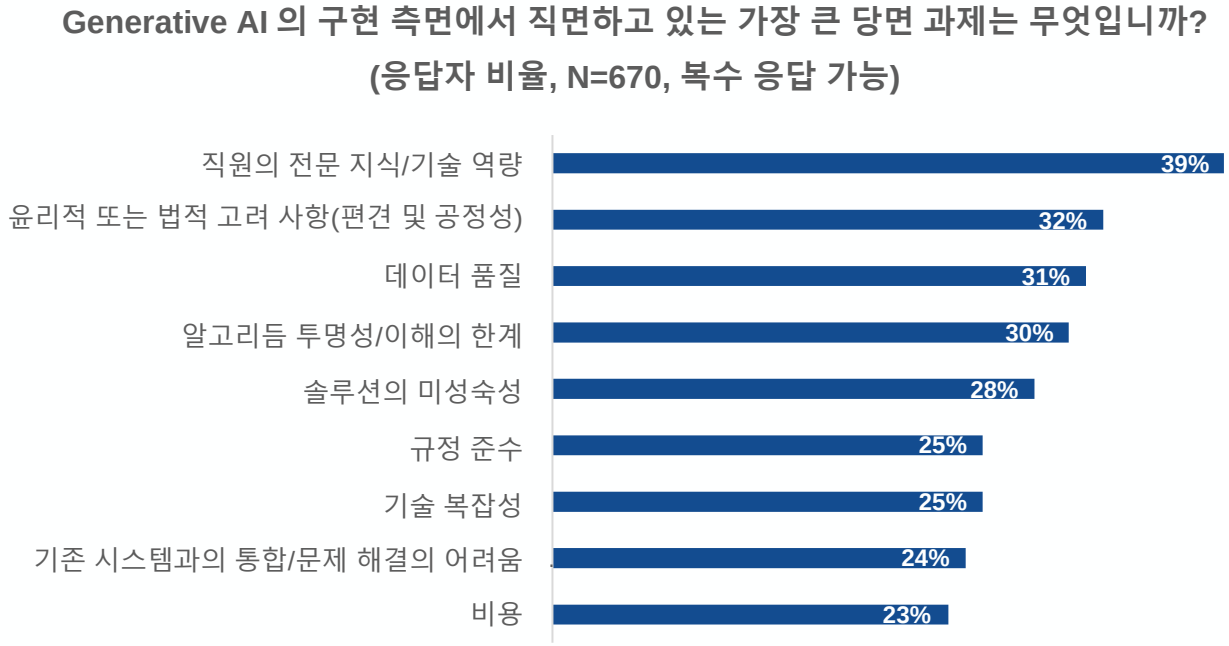
간단히 기존 기술을 교체하는 것이라 할지라도, 새로운 기술을 구축하는 것은 IT 팀에 어려운 일일 수 있다. 특히 완전히 새로운 기술이나 아키텍처는 구축하기가 훨씬 더 어려울 수 있다. 안타깝게도 GenAI에는 새로운 아키텍처가 필요하며, 특히 백엔드 GPU 환경에 새로운 컴퓨팅, 스토리지, 네트워크 인프라스트럭처가 필요하다. 이를 위해 더 많은 인프라스트럭처가 필요할 뿐만 아니라, 무엇보다 GPU 클러스터 전반의 방대한 연결 요구 사항을 수용할 수 있도록 신중하게 설계된 시스템이 필요하다. 400GbE 업링크를 사용하는 일반적인 50GbE(Gigabit Ethernet) 또는 100GbE ToR(Top-of-Rack) 연결은 대규모 언어 모델에 상당한 정체와 지연을 유발하고 전체 이니셔티브를 위험에 빠뜨릴 수 있다.

조직이 Generative AI 솔루션을 구현할 때 직면하는 가장 큰 당면 과제에 대해 물었을 때, 설문조사 응답자들이 꼽은 문제로 데이터 품질, 윤리적 고려 사항, 투명성과 관련된 다양한 당면 과제들이 있었지만, 그중에서도 직원의 전문 지식과 기술 역량 부족, 기술 복잡성, 기존 또는 레거시 시스템과의 통합 불가능, 비용 등이 두드러졌다(그림 2 참조).[5]

[4] 같은 문서 참조.

[5] 출처: Enterprise Strategy Group Complete Survey Results, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, 2023년 8월.

그림 2. GenAI의 주요 당면 과제



출처: Enterprise Strategy Group, TechTarget, Inc. 산하 사업부

특히 Generative AI와 같은 새로운 기술을 도입할 때 기술 역량과 전문 지식의 부족이 가장 큰 당면 과제라는 사실은 놀랄 일이 아니다. 대부분의 조직은 대규모 GenAI 인프라스트럭처, 특히 성능 집약적인 백엔드 환경을 평가, 설계, 구현하는 데 필요한 기술 역량을 갖춘 리소스가 없을 것이다.

기술 복잡성도 GenAI 구축에 영향을 미칠 수 있다. 일부 솔루션은 일반적으로 HPC(High Performance Computing) 환경을 위해 예약된 InfiniBand 네트워크와 같은 독점 기술을 활용하기 때문이다. 그 결과, 적절한 기술 역량을 갖춘 리소스의 수가 제한되어 있다. 특히 이더넷 네트워크를 표준으로 채택한 기업과 하이퍼스케일러의 경우는 더욱 그렇다. 독점 솔루션에는 추가 기술 역량과 하드웨어 및 소프트웨어가 필요하므로 기존 모니터링 또는 오케스트레이션 플랫폼에 통합하기가 더 어려울 수도 있다. 독점 솔루션을 활용할 때 고려해야 할 또 다른 사항은 리드 타임이다. 지난 몇 년 동안 공급망과 관련해 발생한 복잡성을 감안할 때, 조직은 단일 공급업체의 솔루션을 선택하기를 꺼릴 수 있다.

이러한 당면 과제로 인해 조직은 새로운 GenAI 솔루션을 구현하는 데 수반되는 높은 비용으로 인해서도 어려움을 겪고 있다. 특히 확장할 때 특정 공급업체에 종속되는 독점 솔루션이라면 더욱 그렇다. 레퍼런스 디자인과 아키텍처가 부족하면 솔루션을 평가하고 설계하는 데 시간이 상당히 오래 걸릴 수 있다.

조직에 필요한 개방적이고 강력한 GenAI 인프라스트럭처

이러한 점을 고려할 때, 조직은 GenAI 인프라스트럭처의 구축을 가속화하는 데 도움이 되는 개방형 솔루션을 찾아야 한다. 웹 기반 인터페이스를 통한 사용자 상호 작용이 가능하고, 편리한 사용과 액세스에 중점을 둔 새로운 프런트엔드 환경을 구축해야 한다. 백엔드 인프라스트럭처 역시 기존 환경이나 HPC 환경과는 크게 다르다. 방대한 양의 데이터를 소비할 수 있는 GPU 클러스터를 기반으로 하는 LMM(Large Language Model)을 지원해야 한다. 이러한 백엔드 인프라스트럭처 환경은 성공적인 GenAI 프로젝트에 매우 중요하다.

이상적으로 볼 때, 이러한 솔루션은 다음과 같아야 한다.

- **포괄적.** GenAI 솔루션을 구축하려는 조직은 도입을 가속화하기 위해 프런트엔드 환경과 백엔드 환경 모두를 위한 완전한 솔루션이 필요하다. 이러한 솔루션에는 두 환경 모두에 적합한 컴퓨팅(GPU 클러스터 포함), 스토리지 및 네트워킹이 포함된다. 인프라스트럭처 외에도 이러한 솔루션에는 초기 구성과 지속적인 관리뿐만 아니라 패브릭 최적화와 성능의 정밀 조정을 위한 포괄적인 자동화 및 모니터링 툴이 필요하다.
- **고성능.** 네트워크에서 고성능이란 안정적인 제공, 높은 대역폭, 짧은 레이턴시를 지원하는 비차단 패브릭을 구축하는 것을 의미한다. 이것이 바로 Linux Foundation에서 Joint Development Foundation의 일부로 UEC(Ultra Ethernet Consortium)를 설립한 이유이다. 이 컨소시엄은 RoCE v2 프로토콜 등을 통한 한 차원 높은 성능, 확장성, 안정성, 상호 운용성을 통해 AI 환경을 강화하는 이더넷 사양 및 소프트웨어 API 개발에 대한 업계 전반의 협력을 위해 기업을 한데 모으고 있다.[6]
- **사전 테스트와 검증.** 이러한 새로운 GenAI 환경의 도입을 가속화하기 위해, 테스트를 거쳐 효과적으로 작동하는 것이 입증된 포괄적인 솔루션을 배포하는 능력을 갖춘다면 흔히 발생하는 구축 관련 문제를 방지하는 데 도움이 될 수 있다. 이러한 솔루션을 사용하면 연구, 분석 및 설계 시간이 많이 필요하지 않으므로 조직은 GenAI 환경의 목표와 실제 가치를 더 빠르게 달성할 수 있다.
- **개방성 및 확장성.** 여기에는 독점 네트워크 기술이 아닌 상용 표준 규격화 칩과 이더넷 패브릭을 활용하는 것이 포함된다. GenAI 환경에는 가능한 한 높은 네트워크 성능이 필요하지만, 독점이 아닌 개방형 표준으로 이를 얻어야 한다. 이를 위해 UEC는 이더넷이 GenAI 환경에서 중요한 역할을 할 수 있도록 유지할 것이다. 또한 조직은 SONiC(Software for Open Networking in the Cloud)과 같이 상업적으로 이용 가능한 오픈 소스 네트워크 운영 체제도 활용할 수 있다. 참고로 SONiC과 UEC 프로젝트는 모두 Linux Foundation에서 호스팅하므로 업계 협업과 혁신이 용이하다.

[6] Ultra Ethernet Consortium.

Enterprise Strategy Group의 연구에 따르면 온프레미스 데이터 센터를 현대화하려는 조직은 온프레미스에서 하이퍼스케일 솔루션을 활용하는 것을 최우선 과제로 꼽았다.[7]

- **전문 서비스로 보강.** 관련 전문 지식과 경험을 제공할 수 있는 파트너의 도움을 받으면 GenAI 솔루션의 가치 실현 시간을 단축하는 능력이 배가될 수 있다. 이러한 도움에는 적절한 평가를 수행하고, 환경을 설계하고, 적시에 솔루션을 구현하는 능력이 포함된다. 여기에는 완전한 매니지드 서비스와 기술 청사진 또는 검증된 설계도 포함될 수 있다.
- **확장성.** 대부분의 조직이 아직 GenAI의 시작 단계이기 때문에 첫 구축의 규모는 작을 수 있지만, 앞으로 늘어날 요구 사항을 수용하려면 확장해야 한다. 따라서 이러한 요구를 지원하기 위해 GenAI 인프라스트럭처, 특히 네트워크 환경을 확장할 수 있는 능력이 절실하다.
- **에너지 효율.** GPU 기반 솔루션에는 엄청난 전력이 필요하다. 따라서 조직은 소비되는 전력량을 줄이기 위해 가능한 모든 조치를 취해야 하며 처리량 대비 전력 비율을 최적화하는 최신 세대의 실리콘 기술을 사용해야 한다. 고속 스위치는 더 적은 랙 공간, 전력, 케이블 연결을 사용하므로 보다 비용 효율적이고 환경 친화적인 솔루션을 제공할 수 있다. 전력을 줄이는 것 외에 환경 보호 및 지속 가능한 발전 보고서를 제공할 수도 있으므로 운영 및 관리 팀에도 도움이 될 것이다.
- **소프트웨어 기반.** 소프트웨어에 초점을 맞추면 혁신 속도가 빨라진다. 특히 개방형 환경에서 개발된 소프트웨어의 경우 단일 공급업체가 아니라 수십 개의 조직이 혁신에 기여할 수 있기 때문에 더욱 그렇다.

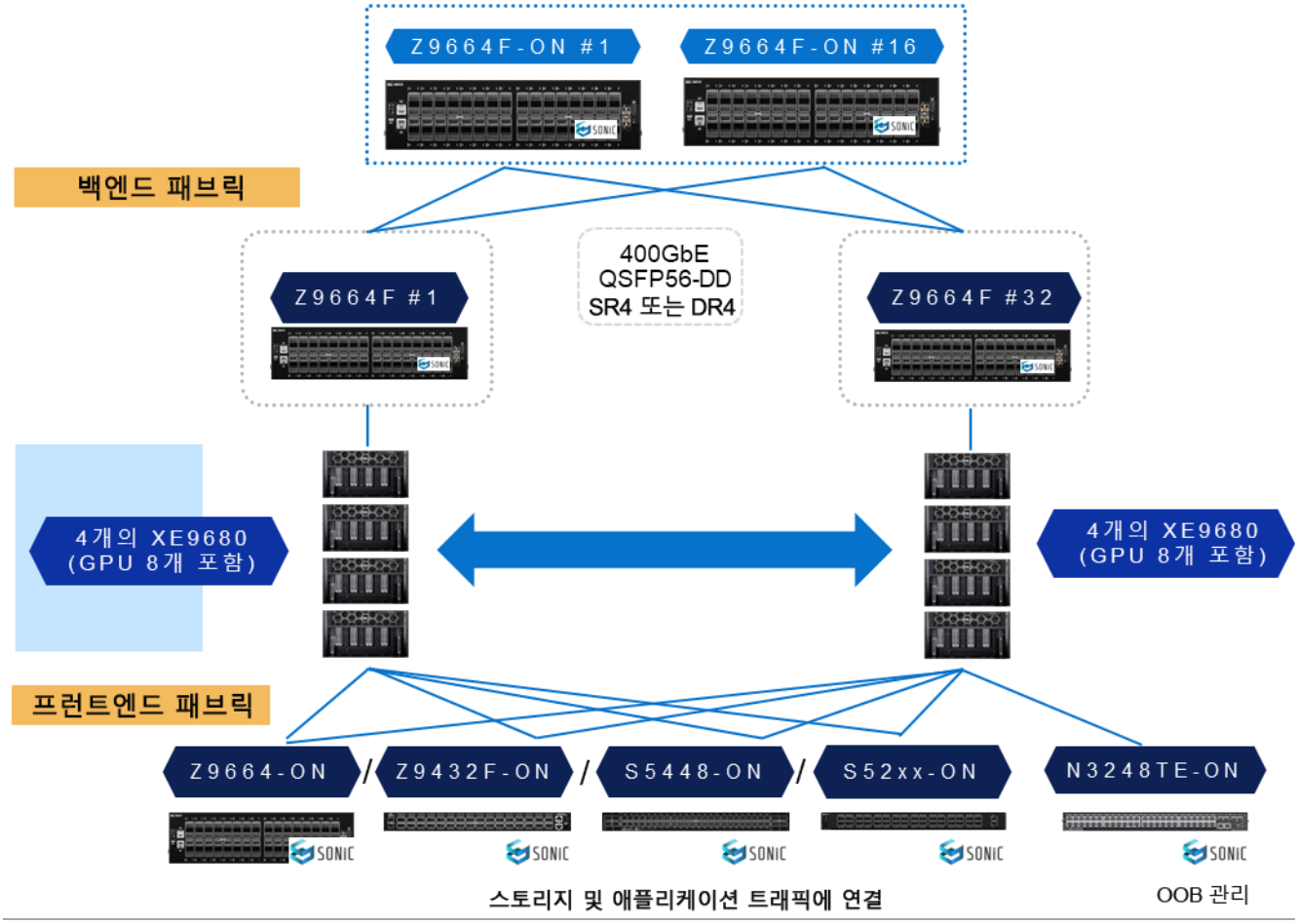
GenAI를 위한 개방형 이더넷 기반 솔루션을 제공하는 Dell Technologies

Dell Technologies는 수년 동안 AI, 모델링 및 HPC 환경을 위한 포괄적이고 개방적인 인프라스트럭처 솔루션을 제공해 왔다. 그리고 이러한 경험을 활용하여 컴퓨팅, 스토리지 및 네트워킹을 포함하는 프런트엔드(애플리케이션 트래픽, 스토리지 액세스, 일반 네트워크) 환경과 백엔드(GPU 패브릭) 환경 모두를 위한 GenAI 인프라스트럭처 솔루션을 지원하고 있다.

고성능 GenAI 솔루션을 가능하게 하는 핵심 요소 중 하나는 그림 3과 같이 검증된 오픈 AI 네트워크 패브릭이다.

[7] 출처: Enterprise Strategy Group Research Report, *2023 Technology Spending Intentions Survey*, 2022년 11월.

그림 3. 포괄적인 AI 네트워크 패브릭 솔루션



출처: Dell Technologies.

Dell Technologies GenAI 솔루션에는 다음이 포함되어 있다.

- **모듈형 컴퓨팅 시스템.** Dell PowerEdge XE 서버를 기반으로 하며 AI, 모델링 및 HPC 시장에서 기업이 쌓은 경험을 바탕으로 제작된 이 서버는 가속에 최적화되어 있다. 공기 냉각 또는 액체 냉각 옵션과 여러 GPU를 지원하며 LLM의 추론 또는 교육에 중점을 둔 Dell은 GenAI 컴퓨팅 요구 사항에 적합한 폼 팩터와 고성능 솔루션을 제공한다. 컴퓨팅 환경은 GenAI를 위한 검증된 설계 및 아키텍처 솔루션의 일부이다.
- **AI 중심 스토리지.** Dell은 PowerScale, Elastic Cloud Storage, ObjectScale 솔루션을 포함하여 워크로드 요구 사항에 따라 선택할 수 있는 다양한 스토리지 옵션을 제공한다. 이더넷 기반 PowerScale OneFS 스토리지는 읽기 및 쓰기 스트리밍이 가능하므로 AI 워크로드의 데이터에 빠르게 액세스하고 AI 모델링 기능을 향상한다. Dell은 PowerScale이 GPU 워크로드를 실행하는 1,000곳 이상의 고객을 대상으로 현장 테스트를 거쳤다고

밝혔다. 그 결과, 이러한 경험을 기반으로 한 수많은 Dell Validated Design 솔루션이 개발되었다. 또한 이렇게 다양한 옵션 모두 Energy Star 인증을 받았다.

- **차세대 이더넷 패브릭.** Dell PowerSwitch를 중심으로 Broadcom Tomahawk 4와 같은 차세대 실리콘을 사용하는 이 오픈 네트워크 하드웨어는 공유 패킷 버퍼링으로 최대 51.2Tbps의 성능을 발휘할 수 있다. PowerSwitch Z Series라는 이름으로 상업적으로 제공되는 Z9664F-ON 64포트 스위치와 Z9432F-ON 32포트 스위치는 수천 개의 노드를 지원하도록 확장할 수 있다. 또한 Dell Technologies는 UEC의 구성원으로서 GenAI 환경에 이더넷을 활용하는 능력을 확대하는 데 기여할 것이다.
- **소프트웨어 기반 아키텍처.** Dell Technologies는 GenAI 환경에서 네트워크 운영 체제, 오케스트레이션 및 모니터링을 위한 오픈 네트워킹 솔루션을 제공하기 위해 최선을 다하고 있다. 네트워크 운영 체제의 경우 Dell Technologies는 SONiC을 수용하고 강화하여 대규모 기업에 필요한 글로벌 지원, 확장성 및 기능을 제공한다. 최신 Enterprise SONiC Distribution by Dell Technologies(버전 4.2)는 RoCE v2(RDMA over Converged Ethernet version 2), 향상된 해싱, 컷스루 스위칭을 포함하여 AI 환경에 대한 고급 지원을 제공한다. 이후 버전 4.3 릴리스에서는 로드 밸런싱과 매핑이 향상되었다. 모든 SONiC 릴리스는 Z Series 포트폴리오 전반에 걸쳐 테스트와 검증을 거쳤으며 Dell의 타사 애플리케이션 파트너 생태계를 대상으로 한 테스트도 거쳤다.
- **도입과 최적화를 가속화하는 서비스 제공.** Dell Technologies는 24/7 글로벌 지원 외에도 검증된 경험을 지닌 전문 서비스 전문가를 통해 조직이 포괄적인 GenAI 솔루션을 적절하게 평가, 설계 및 구현할 수 있도록 돕는다. 이들은 네트워크뿐만 아니라 컴퓨팅과 스토리지 분야도 이해할 수 있으므로 설계 프로세스가 빨라지고 호환성 문제가 발생할 위험이 줄어든다. 이러한 검증된 설계는 추론과 모델 맞춤화를 모두 지원하며, GenAI 파이프라인을 위한 데이터 준비와 수집을 지원하는 서비스도 있다. Dell은 이러한 AI 환경을 운영하기 위한 매니지드 서비스도 제공한다.
- **지속 가능성에 집중.** GenAI 환경을 대규모로 구축하려면 상당한 전력 자원이 필요하다. 브레이크아웃 모드에서 Dell의 고속 스위치는 필요한 랙 공간, 전력 및 케이블 연결이 더 적다. 또한 최신 실리콘 기술을 활용하여 서버, 네트워킹 및 스토리지 솔루션의 에너지 효율성을 최대한 높일 수 있으며 에너지 효율성에 초점을 맞춰 조직에서 비용과 에너지 소비를 줄일 수 있다.

이러한 통합을 통해 Dell Technologies는 백엔드 환경과 프런트엔드 환경 모두를 위한 완전한 GenAI 인프라스트럭처 솔루션을 제공할 수 있는 유리한 위치에 올라 있다.

결론

GenAI에 대한 관심과 활동이 급증하면서 조직은 자체 환경에 적합한 솔루션을 평가하고 있다. 그러나 이러한 유명세는 비교적 최근의 일이므로, 대부분의 IT 팀은 적시에 솔루션을 구현할 수 있는 전문 지식이나 경험이 부족하다. 또 한편으로, 새로운 아키텍처와 기술이 필요한 이러한 GenAI 인프라스트럭처는 매우 복잡하다. 신중하게 설계되고 균형 잡힌 시스템이 있어야 하므로 각 구성 요소를 별도로 소싱하여 통합하려고 하면 매우

위험할 수 있다. 따라서 조직이 성공적인 GenAI 환경을 구축하려면 기술 역량 및 긴밀하게 통합된 솔루션을 확보하기 위해 파트너와 전략적으로 협력해야 한다.

게다가 이러한 환경이 확장될 것을 고려하면 독점 기술을 사용하는 포괄적인 솔루션은 지양할 필요가 있다. 개방형 솔루션은 대규모 GenAI 환경을 위한 혁신, 유연성 및 비용 효율성을 제공할 수 있다. 그리고 견고한 환경을 위해서는 이러한 개방형 솔루션이 완전한 테스트와 검증을 거쳤고 잘 지원되는지 확인하는 것도 중요하다.

Dell Technologies는 프런트엔드 환경과 백엔드 환경 모두를 위한 오케스트레이션과 관리를 포함하여 모든 인프라스트럭처와 소프트웨어를 통합하는 완전한 GenAI 솔루션을 제공한다. 오픈 컴퓨팅, 스토리지 및 네트워킹 또한 여기에 통합되며, 매니지드 서비스, 전문 서비스, Dell의 파트너 생태계를 포함하는 검증된 설계 및 아키텍처를 모두 활용할 수 있다. 이렇게 포괄적이지만 모듈형인 솔루션으로 조직은 GenAI 솔루션의 구축과 가치 실현 시간을 가속화하는 동시에 위험을 줄이고 운영 효율성을 높일 수 있다.

Enterprise Strategy Group 소개

TechTarget의 Enterprise Strategy Group은 집중적이고 실용적인 시장 정보, 수요측 조사, 분석가 자문 서비스, GTM 전략 지침, 솔루션 검증, 엔터프라이즈 기술 구매 및 판매를 지원하는 맞춤형 콘텐츠를 제공합니다.

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