Deloitte.

생성형 AI 열풍, 반도체 업계에 순풍... 거품으로 끝나지 않으려면?

Duncan Stewart 딜로이트 캐나다 TMT Research Director 외 3인

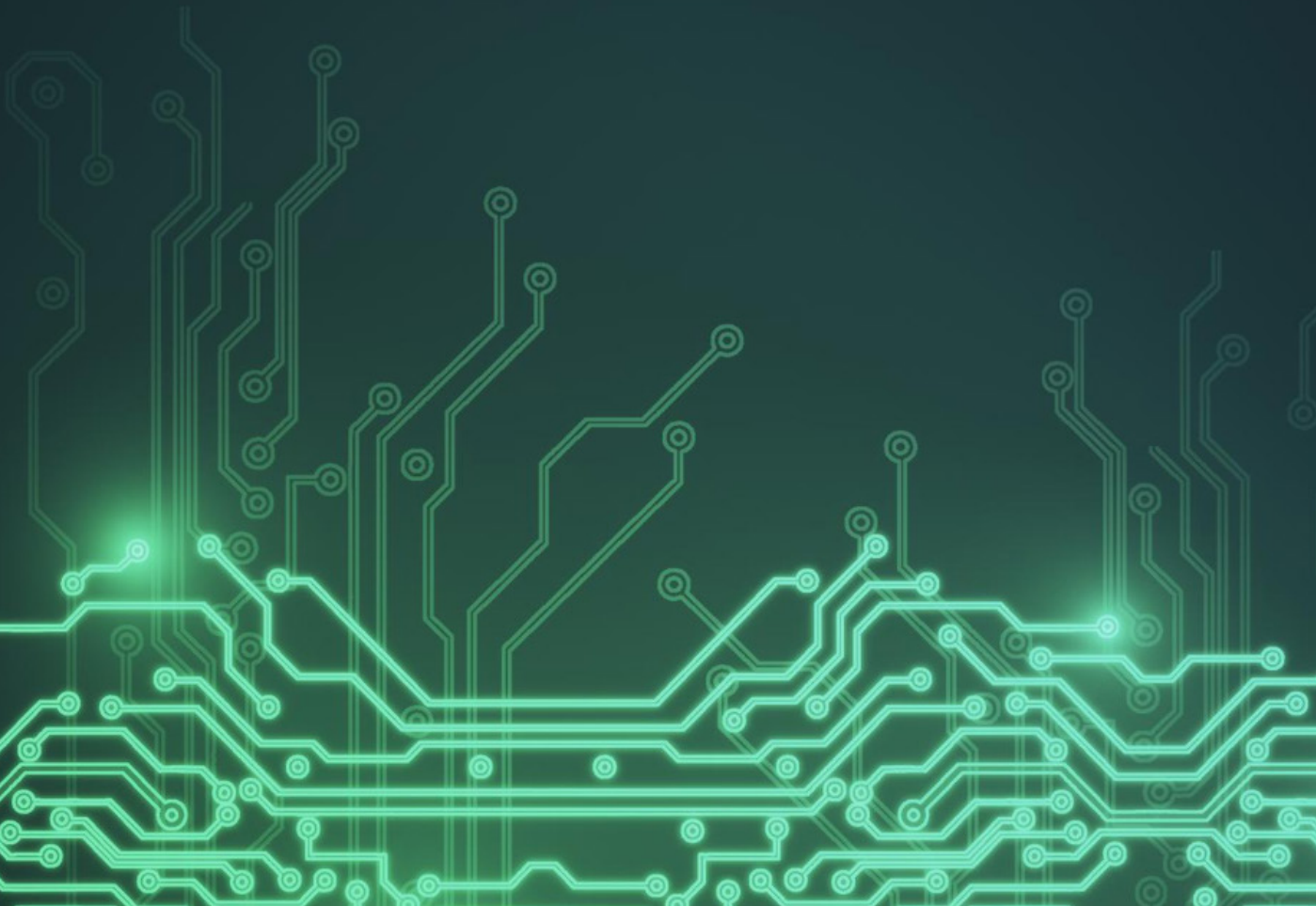
2024년 2월
Deloitte Insights

생성형AI(generative AI) 반도체칩 시장이 빠르게 성장하며, 2027년에는 생성형AI 칩을 포함한 AI 칩이 반도체 전체 매출의 절반을 차지할 것으로 전망된다.

딜로이트는 2024년 생성형AI 전용 반도체칩 시장 규모가 500억 달러를 넘을 것으로 전망한다. 2022년 0달러에 가까웠던 시장이 무섭게 성장해, 2024년 AI 칩 총매출의 2/3를 차지하는 것이다. 이에 따라 2024년에는 글로벌 칩 매출액 5,760억 달러(전망치)에서 AI 칩 총매출이 11%를 차지할 것으로 전망된다.[1] 참고로 최근 제시되는 2027년 AI 칩 시장 규모 전망치는 낮게는 1,100억 달러부터 4,000억 달러까지 범위가 넓다. 물론 보수적인 전망일수록 현실화될 가능성이 크다.[2]

한편으로는 생성형AI 칩 시장의 거품에 대한 우려도 있다. 2023~2024년 막대한 매출이 기대되지만, 기업용 생성형AI 활용사례가 실현되지 않아 2025년 AI 칩 수요가 급감할 수 있다. 2018년과 2021년 암호화폐 채굴 칩 시장의 거품 붕괴와 같은 일이 반복될 수 있는 것이다.[3]

하지만 거품이 붕괴하지 않는다면 보수적인 전망이라 할지라도 AI 칩은 반도체 시장에서 큰 비중을 차지할 것이다. 또한 스마트폰[4] 및 PC[5] 뿐 아니라 AI 칩 시장보다 성숙한 데이터센터 칩 시장의 수요가 부진할 것으로 예상되는 만큼, 반도체 시장 성장에 지금 당장 필요한 순풍 역할을 것이다.



AI 칩 수요 동향

생성형AI는 딥러닝에 신경망을 결합한 머신러닝의 일종으로, 최근 수년간 등장한 여타 AI와 비슷한 기제로 운영된다. 하지만 구세대 AI 칩으로는 생성형AI를 운영하기가 힘들다. 너무 느리고 비효율적이며, 설계 방식과 메모리도 적합하지 않기 때문이다.[6] 따라서 주요 반도체 회사들은 생성형AI에 최적화된 칩을 만들고 있다.

2023년 봄 기준 생성형AI에 특화된 첨단 칩 가격은 개당 약 4만 달러에 달했다.[7] 수요량은 백만 개가 넘을 정도였는데, 생산량이 부족해 공급이 부족했다. 첨단 패키징 병목현상이 주요 원인이었다. 이에 따라 수천 개 기업이 생성형AI 서비스 및 소프트웨어를 출시하는데 애를 먹었다.[8] 생성형AI 칩을 생산할 수 있는 기업들 대다수는 여전히 수주 물량을 맞추지 못해 허덕이고 있고, 이러한 수급난은 2024년까지 지속될 것으로 전망된다.[9] 이처럼 수요는 높은데 공급이 달리면 가격이 높아질 수밖에 없다.

지정학적 요인이 생성형AI 칩 수급에 더욱 큰 영향을 미칠 수도 있다. 생성형AI 특화 칩을 생산하려면 전 세계의 첨단 기술이 집약돼야 하는데, 현재로서는 대부분 공정이 아시아에 집중돼 있고 앞으로도 이러한 상황은 지속될 가능성이 크다. 더군다나 이러한 첨단 칩은 미국과 유럽뿐 아니라 여타 아시아 동맹국들의 대중·대러 무역제한 조치에 포함되는 경우가 많아지고 있다.[10] 중국은 생성형AI 데이터세트와 소프트웨어를 자체 개발할 수 있지만, 향후 5년간은 최첨단 AI 프로세싱에 필요한 첨단 칩을 수입 또는 생산하기가 한층 어려워질 것이다. 여러 제재를 뚫고 중국이 칩 생산 능력을 얼마나 첨단화 시켰는지도 불확실하다. 2023년 9월에 중국 반도체회사가 7나노미터(nm) 프로세스 노드를 기반으로 만들어 내놓은 스마트폰용 칩은 첨단 생성형AI 칩에 사용되는 칩보다 용량이 적고 2~3세대 뒤처졌지만, 여러 무역제재 상황을 반영한다면 의외로 선전했다는 서방 전문가들의 평가를 받았다.[11]

생성형AI 칩 관련 핵심 기술 이슈 세 가지

1. 최첨단 생성형AI 하드웨어의 핵심은 다양한 칩과 연결망으로 이뤄진 랙스케일 보드(rack-scale board)라 할 수 있다. 랙스케일 보드는 중앙처리장치(CPU)와 최첨단 프로세스 노드 기반의 대규모 그래픽처리장치(GPU)를 결합한 것인데, 이러한 GPU는 특수 고속 메모리로 특수 패키징 프로세스를 거쳐 양산된다.[12] 예를 들어, 칩으로서는 대규모에 해당하는 800mm^2 이상의 실리콘 칩에 800억 개의 트랜지스터를 탑재하고 2.5D 첨단 패키징이라 불리는 고대역폭메모리(HBM)3 패키징 프로세스를 거쳐 생성형AI를 운영할 수 있는 GPU가 만들어진다.[13] 이 공정은 파운드리에서의 마지막 프로세스 또는 아웃소싱 업체가 실행하는 백엔드(back-end) 공정의 첫 프로세스로 작업할 수 있다.[14]
2. 이러한 생성형AI 가속기 대부분이 배치되는 데이터센터에서는 대량의 데이터를 가능한 한 빠른 속도로 단거리 이동시켜야 하는 경우가 있는데, 이 때 특수 네트워킹 칩이 필요하다.[15] 네트워킹 칩은 생성형AI 애플리케이션에만 사용되는 것은 아니지만, 현재로서는 생성형AI에 가장 많이 사용되며 2024년 수십억 달러의 매출이 기대되는 분야다.[16]
3. 마지막으로, 생성형AI 칩은 보드당 약 10KW가 필요할 정도로 양산 시 에너지 소비량이 막대하다. 또한 여러 개의 칩으로 이뤄져 있어 냉각기가 감당할 수 있는 것보다 많은 열을 낸다. 따라서 데이터센터 액체냉각(liquid cooling) 시장이 2024년 연간 약 25% 성장해 20억~30억 달러 규모에 달할 것으로 전망된다.[17] 이처럼 대용량 전력을 충당하려면 에너지 효율성이 높은 고압 전력 시설의 신설이 필요하다.[18] 이로 인해 소형 업체들을 중심으로 연간 수억 달러 규모의 고압 전력 시장도 형성될 것으로 전망된다.[19]

결론: 생성형AI, 반도체 시장에 순풍 불어주겠지만 여러 변수 가능성

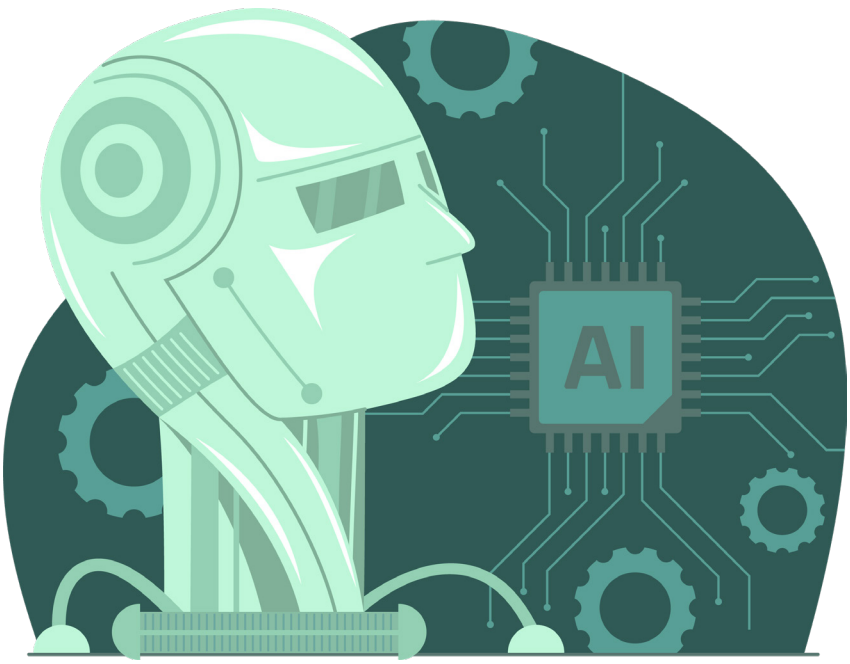
딜로이트는 2024년 반도체 시장에서 생성형AI로 창출되는 매출이 약 500억 달러에 달할 것으로 거의 확신한다. 하지만 현재의 높은 수요와 칩 가격이 공급 증대와 새로운 공급업체의 등장으로 해소된다면 전망은 불확실해진다.

서두에 2027년에 이르면 글로벌 AI 칩 시장 규모가 최대 4,000억 달러에 달할 수 있다는 전망을 소개했다. 이는 꽤 탄탄한 근거가 있는 전망이고 전 세계 반도체 업계도 이러한 전망을 중요하게 여긴다. 하지만 4,000억 달러가 지나치게 낙관적인 전망이라는 근거도 무시할 수 없다.

첫째, 생성형AI GPU 시장은 2023년 여름 기준 단 한 개의 기업이 독점해, 공급이 부족할 수밖에 없었다.[20] 한편 구매자들은 소비자 및 기업용 생성형AI 학습 및 추론에 쓰일 칩이 더 많이 필요할 것이라는 전망에 되도록 많은 칩을 확보하려 사재기에 나섰다.[21] 그 결과 칩 가격이 천정부지로 치솟았다. 하지만 현재 시장을 독점한 기업의 생산량이 늘거나 새로운 경쟁사가 등장한다면 생성형AI 칩 가격은 하락할 가능성이 크다. 그렇게 되면 이들 기업의 매출은 2025년부터 감소할 수 있다.

둘째, 칩을 구매하는 고객사들은 충분한 물량을 공급받지 못할 때 필요보다 많이 주문한다. 예를 들어, 주문한 물량의 25%밖에 공급받지 못할 것이라는 계산에 실제로 필요한 물량은 2만5,000 개인데 10만 개를 주문하는 것이다. 수요량이 7만5,000개 부풀려지는 것이다. 하지만 AI 칩 생산량이 늘어 수요-공급 균형이 회복될 경우, 구매 기업들이 이전대로 주문하면 필요 이상의 칩을 떠안게 되기 때문에 주문량을 줄일 것이다. 과거 반도체 산업의 급격한 주기 변동을 야기한 '채찍 효과'가 발생하는 것이다.[22]

셋째, 현재 생성형AI 학습과 추론은 모두 데이터센터용 생성형AI 칩과 동일한 칩으로 운영된다. 하지만 앞으로 생성형AI 추론의 상당 부분이 엣지 프로세서에서 이뤄질 가능성이 크다.[23] 엣지 프로세서 생성형AI는 데이터센터용보다 성능이 낮은 GPU 및 CPU, 또는 새로운 애플리케이션 특화 집적회로(IC)로도 운영이 가능하다. 따라서 관련 시장에는 기존 생성형AI 칩 제조사뿐 아니라 기존 엣지 프로세싱 칩 회사와 반도체 설계를 하지 않았던 기업들까지 새로운 경쟁사들이 진입할 수 있다.[24] 엣지 프로세서에서 처리되는 생성형AI의 추론이 늘어날수록, 시장 규모가 커지거나 데이터센터용 생성형AI 칩 가격이 하락할 것이다.



마지막으로, 앞서 언급했듯 생성형AI 칩 시장이 2023~2024년 폭발적으로 성장하다가 2025년 거품이 붕괴될 것이라는 우려가 있다. 이는 대세 전망은 아니지만, 반도체 시장의 난폭한 등락 가능성을 익히 알고 있다면 완전히 무시할 수 있는 견해는 아니다.

확실히 전망하기는 어렵지만 AI 칩 공급이 늘고 다각화되는 한편 수

요가 예상보다 저조하고 엣지 프로세서에서 이뤄지는 AI 추론이 늘어나 가격이 하락하면, 2027년 AI 칩 시장 규모는 앞서 언급한 1,100억~4,000억 달러의 하단에 그칠 가능성이 있다. 그렇다 하더라도 2024년 규모(전망치)에서 두 배 성장하는 수준이다.

하지만 시장 규모가 1,000억 달러든 4,000억 달러든 기업들이 AI 칩, 특히 생성형AI 칩을 필요로 할 것이라는 점은 변함이 없다. 또한 안정적이고 신뢰할 수 있는 공급망을 확보해야만 혁신, 경제적 성공, 국가 안보를 사수할 수 있다는 기업들의 인식도 변하지 않을 것이다.

이 대목에서 미국과 유럽의 해결과제가 이슈로 등장한다. 미국과 유럽에서 다수의 반도체회사들이 AI 및 생성형AI 운영에 필요한 첨단 CPU와 GPU를 생산할 수 있는 노드 제조시설을 구축하고 있지만,[25] 프론트엔드(front-end)와 백엔드(back-end)를 통틀어 패키징 역량이 현재로서는 부족하다.[26] 뿐만 아니라 AI 칩 필수인 HBM이나 HBM3e 공장도 기존으로는 터무니없이 부족한데 신설 계획도 충분치 않다.[27] 따라서 미국과 유럽은 AI 칩을 자체 생산할 수는 있어도, 결국 HBM3 메모리 탑재와 첨단 패키징 공정을 위해 한국, 대만, 동남아시아 등 아시아에 의존할 수밖에 없는 실정이다.

유럽반도체법(European Chips Act)과 미국 반도체 및 과학법(CHIPS and Science Act) 모두 첨단 패키징 및 메모리 기술 개발과 생산능력 강화를 위한 예산을 책정했으나,[28] 이것만으로 생성형AI 칩 패키징의 자급자족이 가능해질지는 불확실하다.

생성형AI 칩 시장의 성장에 있어 또다른 변수는 중국이다. 현재 미국을 위시해 네덜란드와 일본 등이 대중 수출제한 조치를 발동해, 생성형AI 칩을 포함한 모든 종류의 첨단 노드 기반 칩의 대중 수출뿐 아니라 중국에 기술 노하우를 전수하는 것까지 제한하고 있다.[29] 향후 미국 등의 추가 수출제한 조치로 비(非)첨단 칩까지 수입할 수 없게 될 상황에 대비해,[30] 주요 중국 인터넷 기업들은 2023년 8월에 50억 달러 어치의 생성형AI 칩을 주문한 바 있다.[31]

생성형AI가 2027년에도 지금처럼 혁신, 경제성장, 국가안보에 중요한 기술로 간주되고 중국이 계속 첨단 AI 칩을 수입하지도 못하고 개발에 필요한 툴도 얻지 못한다면, 중국은 AI 칩 양산에 필요한 원재료의 수출 제한이라는 강수를 둘 수 있다. 그렇게 되면 전 세계 AI 칩 생산이 큰 차질을 빚게 되고 글로벌 경제성장에도 막대한 부정적 영향을 미칠 수 있다.

주석

1. World Semiconductor Trade Statistics, "WSTS Semiconductor Market Forecast Spring 2023," May 2023.
2. Analysis based on data sourced from multiple publicly available sources: Martin Baccardax, "Nvidia jumps higher as Mizuho analysts see $300 billion AI chip potential," The Street, July 24, 2023; Patrick Seitz, "Intel on track with data center chip lineup, touts play in artificial intelligence," Investor's Business Daily, March 30, 2023; World Semiconductor Trade Statistics, "WSTS Semiconductor Market Forecast Spring 2023;" Deborah Yao, "Analysts' take: Nvidia widens its total addressable market," AI Business, June 1, 2023.
3. The Economist, "Crypto-miners are probably to blame for the graphics-chip shortage," June 19, 2021.
4. International Data Corporation (IDC), "Weak consumer demand continues to delay a recovery for the smartphone market, according to IDC," press release, May 31, 2023.
5. IDC, "PC and Tablet market face further decline before a rebound in 2024, according to IDC," press release, June 13, 2023.
6. Samuel K. Moore, "Nvidia's Next GPU Shows That Transformers Are Transforming AI," IEE Spectrum, April 8, 2023.
7. Kif Leswing, "Nvidia's top A.I. chips are selling for more than $40,000 on eBay," CNBC, April 14, 2023.
8. GPU Utils, "Nvidia H100 GPUs: supply and demand," July 2023 (updated August 2023), accessed September 15, 2023.
9. Lucas Mearian, "Chip industry strains to meet AI-fueled demands — will smaller LLMs help?," Computerworld, September 28, 2023.
10. Rita Liao, "China's AI firms might further lose chip access in new US ban," TechCrunch, June 28, 2023.
11. Jeff Pau, "SMIC bypasses US curbs to make 7nm chips," Asia Times, September 5, 2023.
12. Dylan Patel, Myron Xie, Gerald Wong, and George Cozma, "AI Capacity Constraints - CoWoS and HBM Supply Chain," Semi Analysis, July 5, 2023.
13. Ibid.
14. Semiconductor Engineering, "Advanced Packaging," accessed November 14, 2023.
15. Brian T. Horowitz, "AI Workloads Spur Competition in Networking Chips," Network Computing, July 13, 2023.
16. Deloitte analysis of AI networking chip market.
17. Deane Dray, Jonathan Atkin, et al., RBC Imagine: Datacenter Liquid Cooling Market Overview, June 21, 2023.
18. Steve Taranovich, "Data centers feel the power density pinch," Electronic Design, August 6, 2021.
19. Dylan Patel, Myron Xie, Gerald Wong, George Cozma, "Energizing AI: Power delivery competition heats up Vicor, MPS, Delta, ADI, Renesas," Semi Analysis, August 2, 2023.
20. Dylan Patel, Myron Xie, Gerald Wong, and George Cozma, "AI Capacity Constraints - CoWoS and HBM Supply Chain."
21. Ibid.
22. Chris Richard, Dan Hamling, Duncan Stewart, and Karthik Ramachandran, Five fixes for the semiconductor chip shortage, Deloitte Insights, December 6, 2021.
23. Lucas Mearian, "Chip industry strains to meet AI-fueled demands — will smaller LLMs help?"
24. Ibid.
25. Michelle Adams, "Where Are All the New Semiconductor Fabs in North America & Europe?" Z2Data, September 12, 2023.
26. Duncan Stewart, Karthik Ramachandran and Brandon Kulik, Chipping in to boost production: US and Europe move toward greater self-sufficiency and resilient supply chains, Deloitte Insights, April 24, 2023.
27. Anton Shilov, "Memory makers on track to double HBM output in 2023," AnandTech, August 9, 2023.
28. Sheryl Miles, "CHIPS Act implementation requires strong focus on 'Advanced Packaging'," Electronic Specifier, October 11, 2022.
29. Anirban Ghoshal, "US wins support from Japan and Netherlands to clip China's chip industry," COMPUTERWORLD, January 30, 2023.
30. Andrew Ross Sorkin, Ravi Mattu, Bernhard Warner, Sarah Kessler, Michael J. de la Merced, and Lauren Hersch, "The A.I. chips war could heat up this summer," The New York Times, June 28, 2023.
31. Kanjyik Ghosh and Stephen Nellis, "China's internet giants order $5 bln of Nvidia chips to power AI ambitions -FT," Reuters, August 10, 2023.

딜로이트 첨단기술, 미디어 및 통신 산업 전문 리더

딜로이트 첨단기술, 미디어 및 통신 산업 전문팀은 빠르게 발전하는 산업 환경 속에서 고객들의 전략적 과제들을 해결할 수 있는 최상의 서비스 경험을 제공합니다. 딜로이트 첨단기술, 미디어 및 통신 산업 전문팀은 국내외 기업의 전략수립, 회계감사, 재무자문, IT 시스템 구축 등 다양한 서비스 경험을 보유한 우수 전문인력으로 구성되어 있습니다.

Contact

김우성 파트너
Technology Strategy & Transformation리더 | 딜로이트컨설팅
Tel: 02 6099 4670
Email: wooskim@deloitte.com

안상혁 파트너
디지털부문 리더/금융산업 총괄리더 I 딜로이트 컨설팅
Tel: 02 6676 3625
Email: sanghyan@deloitte.com

박지숙 파트너
금융 IT, 오퍼레이션 리더 I 딜로이트 컨설팅
Tel: 02 6676 3722
Email: jisukpark@deloitte.com

장지영 파트너
Tech Strategy 부문 파트너 I 딜로이트 컨설팅
Tel: 02 6676 3956
Email: jiyojang@deloitte.com

강기식 파트너
Lead Architect I 딜로이트 컨설팅
Tel: 02 6676 2039
Email: gikang@deloitte.com

주형열 파트너
반도체 CoE 리더 | 딜로이트컨설팅
Tel: 02 6676 3750
Email: hjoo@deloitte.com

최호계 파트너
Technology Sector 리더 | 감사본부
Tel: 02 6676 3227
Email: hogchoi@deloitte.com

박형곤 파트너
TME Sector 리더 | 딜로이트컨설팅
Tel: 02 6676 3684
Email: hypark@deloitte.com

조명수 파트너
Digital Finance & Operation 리더
Tel: 02 6676 2954
Email: mjo@deloitte.com

박권덕 파트너
TME Sector 리더 | 딜로이트 컨설팅
Tel: 02 6676 3567
Email: gwapark@deloitte.com

앱스토어, 구글플레이/카카오톡에서 '딜로이트 인사이트'를 검색해보세요.
더욱 다양한 소식을 만나보실 수 있습니다.

Deloitte.
Insights

성장전략본부 리더
손재호 Partner
jaehoson@deloitte.com

딜로이트 인사이트 리더
정동섭 Partner
dongjeong@deloitte.com

연구원
김선미 Manager
seonmikim@deloitte.com

디자이너
박주리 Consultant
jooripark@deloitte.com

Contact us
krinsightsend@deloitte.com