고성능컴퓨팅특론 중간고사 대체과제

202001488 김문기

Architecting Waferscale Processors - A GPU Case Study

> Saptadeep Pal, Daniel Petrisko, Matthew Tomei, Puneet Gupta, Subramanian S. Iyer, and Rakesh Kumar

Published in: 2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)

communication overheads 가 증가함에 따라 컴퓨터 시스템 확장에 문제를 겪고 있습니다. 논문에 따르면 “One approach to dramatically reduce communication overheads is waferscale processing.” 라고 합니다. 이를 통해 현재 방식에 문제가 있음을 알 수 있습니다. communication overheads 가 어디에서 크게 발생하는지를 안다면 그 부분을 개선했을 때 가장 큰 속도 향상이 있을 것입니다.

waferscale perocessor를 GPU case 에서는 사용하지 않았던 이유는 수율 문제입니다. (“Trilogy Systems. ” https://en.wikipedia.org/wiki/Trilogy Systems, (accessed Nov 20, 2017).) 역사상 가장 큰 지원을 받고 설립된 Trilogy 가 실패한 이유 또한 waferscale processor 에서 이 yield problem 을 피하지 못했던 것을 주요인으로 뽑고 있습니다. 하지만 Silicon-Interconnection Fabric (A. Bajwa and S. Jangam and S. Pal and N. Marathe and T. Bai and T. Fukushima and M. Goorsky and S. S. Iyer, “Heterogeneous Integration at Fine Pitch (≤10µm) Using Thermal Compression Bonding,” in 2017 IEEE 67th Electronic Components and Technology Conference (ECTC), pp. 1276–1284, May 2017.)기술이 등장한 이래로 waferscale processor 는 다시 살펴볼 필요가 있습니다.

conventional integration technology와 논문에서 사용한 integration technology 의 차이점은 아래와 같습니다.

- 기존의 방식
  - single processor die의 최대 크기는 수율에 의해 결정되는 싱글 웨이퍼에서 동시에 제조된다.
  - 제조 후 웨이퍼는 개별 프로세서 다이로 절단 되어 패키징된다.
  - PCB에서 패키징된 프로세서를 연결하는 IO link를 사용하여 병렬 시스템에 통합된다.
- 새로운 방식
  - waferscale processor에서 wafer는 processor다. 즉, monolithic processor가 전체 웨이퍼만큼 크게 설계되거나 프로세서 세트가 웨이퍼와 프로세서 다이에 계속 상주하도록 설계되었다.
  - processor die는 저렴한 on-wafer interconnect를 사용하여 웨이퍼 자체에 연결된다.

두 방식에서 오는 차이점으로 기존에 방식이 어떤 단점을 가지고 있는지 짐작할 수 있습니다. 이 논문에서는 communication overheads를 줄이기 위한 방법으로 waferscale processor 를 제안합니다. 즉, main idea는 PCB 상에서의 비효율적 link를 없애기 위함이라 볼 수 있습니다.

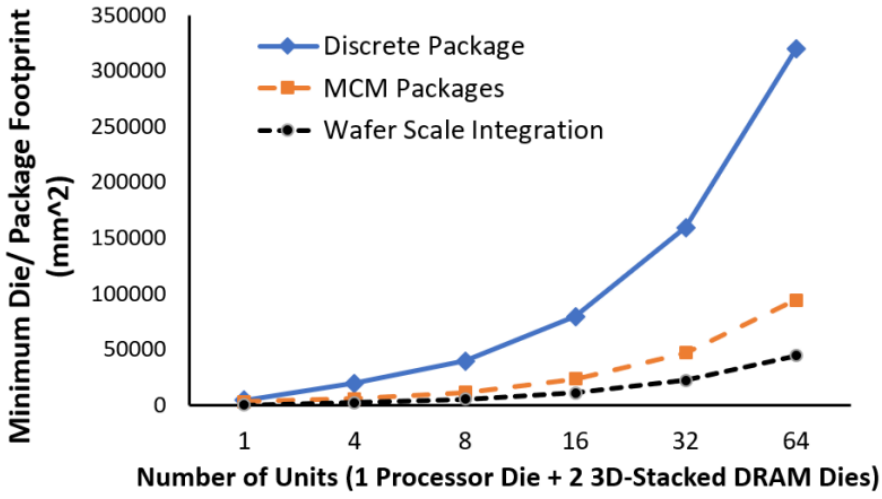

Fig. 1: Minimum die/package footprint for different integration schemes are shown for increasing number of processor dies per system

그림 1은 여러 시나리오에서 computing die의 총 면적 풋프린트를 보여줍니다(Saptadeep Pal, Daniel Petrisko, Matthew Tomei, Puneet Gupta, Subramanian S. Iyer, and Rakesh Kumar. Architecting Waferscale Processors - A GPU Case Study. In HPCA 2019.). 논문에 설명은 아래와 같습니다. “each die is placed in a discrete package, 4 units (each unit consists of a processor die and two 3D-stacked DRAM dies) inside an multi-chip module (MCM) package and, waferscale integration.”

discrete package의 경우 당연한 결과를 보입니다. 각각의 다이가 연결된 곳에서 overheads가 너무 크기 때문입니다. 하지만 MCM 패키지 외부 다이 연결에서 PCB trace 와 PCB 간 링크에서 얼마나 많은 communication overheads가 발생하는지 볼 수 있습니다.

현재로써는 웨이퍼에 수용할 수 있는 GPM에 물리적 문제로 인해 40개 정도로 제한이 있는 상태입니다. 그리고 논문은 현재의 waferscale processor의 한계점을 제시합니다. “We find that waferscale GPUs are area-constrained due to power delivery network overheads, not thermally-constrained.” 이 문제점에 대한 자세한 후속 연구가 필요함을 알립니다.

이전에 연구된 packageless processors based on the Si-IF(S. Pal, D. Petrisko, A. A. Bajwa, P. Gupta, S. S. Iyer, and R. Kumar, “A case for packageless processors,” in 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA), pp. 466–479, Feb 2018.) 에서 패키지 제거로 인해 bandwidth, thermal and area benefits을 얻을 수 있음을 알 수 있었습니다. 하지만 이 논문은 기존과 같은 크기의 싱글 다이 프로세서 시스템뿐만 아니라 wafer에 맞먹는 크기의 GPU system architecture 에 초점을 두고 있습니다.