IEEE EPS TC6 Newsletter

Heterogenous Integration for AI Chips

Yasumitsu Orii and Atsushi Takahashi

Introduction

The amount of compute used in the largest AI training runs has been increasing exponentially with a 3.4-month doubling time. It would be hard to sustain this growth if the AI chip development depends solely on the semiconductor process advancement since Moore's Law has a 2-year doubling period and there is concern it is reaching its physical limit. The question this is how do we improve the computer hardware performance? The Von Neumann bottleneck is always the issue and Heterogenous Integration can provide good solution paths. And in recent years, chiplets, which allow us to achieve efficient high-performance computing better than ever before, have become the focus of the industry in response to the concern that Moore's Law scaling is nearing its end. Another approach to resolve the Von Neumann bottleneck is to implement a neuromorphic device. We will discuss the key interconnection technologies such as high-density substrate, wafer level fan-out and Bridge to support chiplets and Neuromorphic devices.

AI training speed

To meet big data processing demand in IoT and 5G/beyond 5G, AI training like deep learning became the most important algorithms that require a 3 to 4 months FLOPS doubling rate. It is an extremely higher rate much faster than that supported by Moore's Law doubling of transistor density every 2 years. This means a variety of innovative hardware technologies must be developed to fill the big gap. AlphaGo which won against humans in the Go game in 2018 showed AI training computing capability clearly.

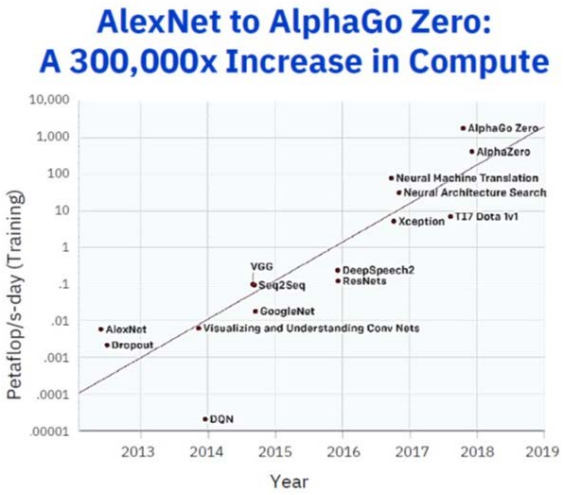

Figure 1: Alexnet to AlphaGo Zero［1］

AI Chip Demand

The AI chip market is expected to grow at more than 35% CAGR in 2020 - 2026 driven by need in the cognitive computing era. The AI accelerator is getting to be a mainstream device in high performance cloud computing at the data center. Inference AI chip and natural language processer for edge computing involving smart phone, autonomous car and medical are also gaining clear presence in IoT and 5G/beyond 5G. The AI chips market is anticipated to get over US$70 billion accounting for about 15% of the semiconductor market share in 2026. AI algorithms based on matrix multiply accumulate are compatible with GPU base architecture. Since low power, low latency and high FLOPS are crucial for AI chips, FPGA based architecture products are being developed.

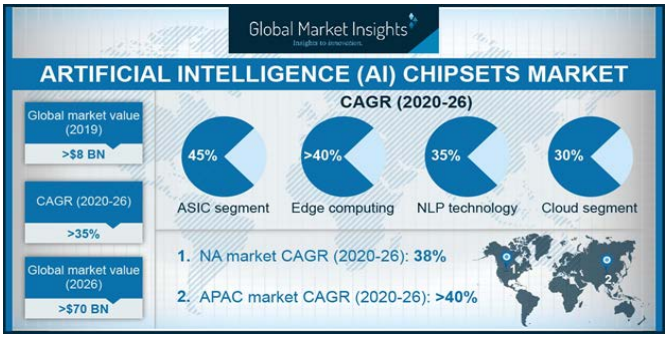

Figure 2: ARTIFICIAL INTELLIGENCE CHIPSETS MARKET［2］

The Market for Chiplets

Chiplets are integrated circuit blocks which are similar to Lego blocks. A chiplet maximizes HPC FLOPS at low power, optimizing each IP technology node by dividing the SoC into each IP block to minimize increasing chip cost in advanced technology nodes and maximizing production yield as the alternative (or complementary) solution to Moore's Law. It started with fine build up substrate based on organic interposer in HPC and will be followed by RDL interposer achieving more design rule flexibility and production scalability than fine build up substrate for HBM application on GPU/ASIC/FPGA as well as CPU/GPGPU/ASIC/FPGA chiplets without HBM. In 2030 RDL will dominate in chiplets with about 60% market share.

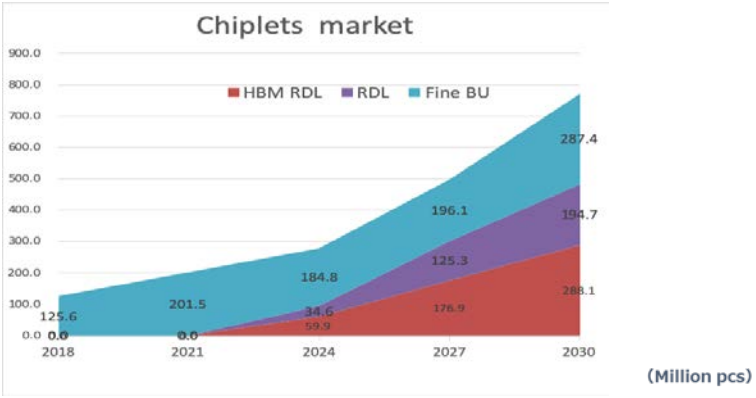

Figure 3: Chiplets market［3］

Neuromorphic device

IBM Almaden developed a new chip, TrueNorth, with a brain-inspired non-Von Neumann computer architecture[4][5]. It consists of one million neurons and 256 million synapses. Built on 28nm process technology, the 5.4 billion transistors chip has an on-chip network of 4,096 neurosynaptic cores as shown in Fig.4. TrueNorth' s power density is 20mW per cm^2 and only consumes 70mW during the real-time operation whereas the power density of a typical central processing unit (CPU) is 50 to 100 W per cm^2 [6]. 3D stacking is a perfect approach for the scaling of neuromorphic devices because of its low power, high performance and miniaturization capabilities. TrueNorth x16 chiplets is equivalent to the brain of frog and TrueNorth x 100 chiplets with high density interconnects equals 100M neurons which is close to a cat brain level application in edge computing. Neuromorphic 3D stacked chiplets are largely expected to mimic the human brain (100B neurons) in future high performance cognitive computing. This is one good example of chip integration for the coming neuromorphic device era.

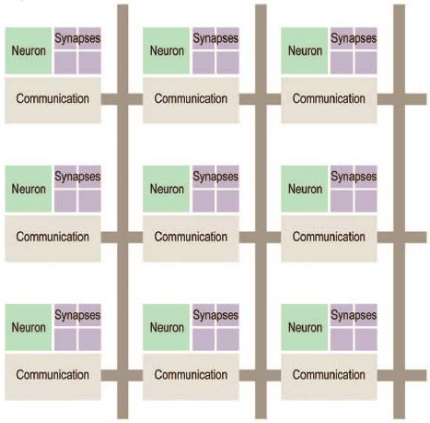

Figure 4: TrueNorth conceptual blueprint of an architecture that, like the brain, tightly integrates memory, computation, and communication[6]

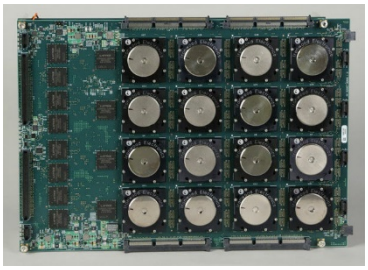

Figure 5: TrueNorth module photo［7］

Interconnection Technologies Discussion

High density interconnect is key technology in chiplets for high performance computing as an alternative to traditional BEOL process. HDI on organic interposer like iTHOP from Shinko is traditional finer pitch type of build-up substrate [8]. Wafer level fan out like InFO or SLIM drives higher signal integrity and power integrity by introducing the RDL process and panel level fan out is expected to extend its scalability. Recently bridge type of interconnect technology like EMIB is expected to achieve not only high density but also improved power delivery networks.

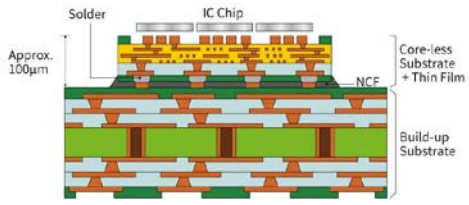

Figure 5: i-THOP［8］

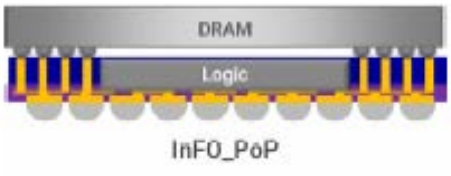

Figure 6: InFO［9］

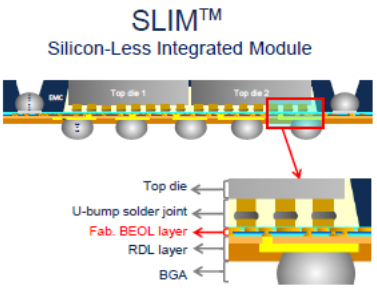

Figure 7: SLIM［10］

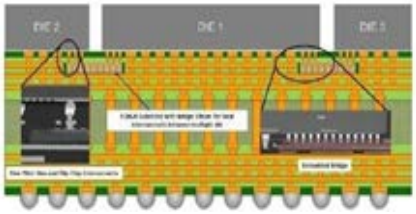

Figure 8: EMIB［11］

[1] D.Amodei, D. Hernandez: https://blog.openai.com/ai-and-compute

[2]Global Market Insights :https://www.gminsights.com/industry-analysis/artificial-intelligence-ai-chipsets-market

[3] Toshihiko Nishio, Chiplets market trend report, March, 2020

[4] Dharmendra Modha, ‘’Introducing a Brain-inspired Computer**,** TrueNorth's neurons to revolutionize system architecture’’ https://www.research.ibm.com/articles/brain-chip.shtml

[5] Yasumitsu Orii, Akihiro Horibe, Kuniaki Sueoka, Keiji Matsumoto, Toyohiro Aoki, Hirokazu Noma,Sayuri Kohara, Keishi Okamoto, Shintaro Yamamichi, Kohji Hosokawa and Hiroyuki Mori, ’’PERSPECTIVE ON REQUIRED PACKAGING TECHNOLOGIES FOR NEUROMORPHIC DEVICES’’,IMAPS2015

[6] Paul A. Merolla, John V.Arthur, Rodrigo Alvarez-Icaza, Andrew S.Cassidy, Jun Sawada, Filipp Akopyan, Bryan L. Jackson, and Dharmendra S. Modha, “A million spiking-neuron integrated circuit with a scalable communication network and interface”,

Science 345, 2014, pp. 668-673.
[7] IBM, https://gigazine.net/gsc_news/en/20150820-ibm-truenorth-for-smartphone/
[8] SHINKO, i-THOP(integrated Thin film High density Organic Package), https://www.shinko.co.jp/product/under-development/i-thop/
[9] TSMC, InFO(Integrated Fan-Out) Wafer Level Packaging, https://3dfabric.tsmc.com/english/dedicatedFoundry/technology/InFO.htm
[10] 3D incites, ‘’Amkor’s SLIM & SWIFT Package Technology’’, http://www.3dincites.com/wp-content/uploads/slim-swift-customer-overview-may-13-2015.pdf
[11]Ravi Mahajan and Robert Sankman and Neha Patel and Dae-Woo kim, ‘’A High Density, High Bandwidth Packaging Interconnect Embedded Multi-die Interconnect Bridge (EMIB)’’, Las Vegas, ECTC 2016