HKSTP 香港科技園

Thought-leadership Series

WHITE PAPER ON AI CHIPS SUMMIT 2019

By Hong Kong Science and Technology Parks Corporation and United Microelectronics Centre (Hong Kong) Limited

Powered by

UMEC
聯合微電子中心(香港)
UNITED MICROELECTRONICS CENTRE (HONG KONG)

CONTENTS

AI Chips Summit 2019

Organisers

Hong Kong Science and Technology Parks Corporation
United Microelectronics Centre (Hong Kong) Limited

Keynote Speakers (Listed Alphabetically)

Prof. Deming Chen	Professor	University of Illinois Urbana-Champaign
Prof. Tim Cheng	Dean of Engineering	The Hong Kong University of Science and Technology
Mr. Nelson Fan	Vice President	ASM Pacific Technology
Dr. Walden Rhines	CEO Emeritus	Mentor, a Siemens Business
Prof. Elyse Rosenbaum	Professor	University of Illinois Urbana-Champaign
Dr. Hayden So	Associate Professor	The University of Hong Kong
Prof. Chi Ying Tsui	Professor	The Hong Kong University of Science and Technology
Prof. Evangeline Young	Professor	The Chinese University of Hong Kong

Panellists (Listed Alphabetically)

Prof. Philip Chan	Deputy President and Provost	The Hong Kong Polytechnic University
Prof. Deming Chen	Professor	University of Illinois Urbana-Champaign
Prof. Tim Cheng	Dean of Engineering	The Hong Kong University of Science and Technology
Mr. Nelson Fan	Vice President	ASM Pacific Technology
Prof. Frank He	Director	The Key Laboratory of Integrated Microsystems, Peking University
Dr. Mei Kei Ieong	CEO	United Microelectronics Centre (Hong Kong) Limited
Mr. Lincoln Lee	PacRim Technical Director	Mentor, a Siemens Business
Mr. Matthew Leung	Director	Hong Kong Research Centre, Huawei
Dr. Walden Rhines	CEO Emeritus	Mentor, a Siemens Business
Prof. Martin Wong	Dean of Engineering	The Chinese University of Hong Kong
Dr. H.L. Yiu	Head of Advanced Manufacturing	Hong Kong Science and Technology Parks Corporation
Mr. Wilson Yu	Board Director	China Resources Microelectronics Ltd.

EXECUTIVE SUMMARY

EXECUTIVE SUMMARY

Artificial Intelligence (AI) will fundamentally reshape the way businesses, governments, and individuals operate on a day-to-day basis. The rapid growth of AI means there is an urgent need for more powerful and energy-efficient AI chip design architecture to maintain the sustainable development of the industry.

AI Chips Summit, co-organised by Hong Kong Science and Technology Parks Corporation and United Microelectronics Centre (Hong Kong) Limited, held on 5 December 2019 at Hong Kong Science Park. The summit aims to explore the latest technologies of AI chips and system, discuss the industry development and foster collaboration between various AI practitioners. Global AI leading experts from the academia and industry are invited to share the latest technologies of AI chips, as well as to discuss the opportunities, challenges and strategies for the industry development. The summit was divided into two sessions. The first session focused on the AI-assisted design tool for AI hardware. Four keynote speakers shared their insights into the electronic design automation (EDA) tools development and introduced their latest research result. A panel discussion was carried out to discuss about the new EDA tools and the opportunities brought to the start-ups. The second session was about heterogeneous integration and hardware-aware design algorithms. Four keynote speakers shared their experience in 3DIC heterogeneous integration. After that, another panel discussion was carried out to discuss about the future development of the microelectronic industry in Hong Kong.

Mr. Albert Wong,
Chief Executive Officer of Hong Kong Science and Technology Parks Corporation

RAPID GROWTH OF AI CHIP AND MICROELECTRONICS IN HONG KONG

AI chips are developing rapidly nowadays. Because of the computational power required and the interconnected demography, the market will grow from US$6 billion now to more than US$90 billion in 2025. This is a good opportunity for the industry and academia to get together to promote the development, according to Mr. Albert Wong, Chief Executive Officer of Hong Kong Science and Technology Parks Corporation.

There are about 80 to 90 companies in Science Park working on microelectronics, which is a great concentration of R&D people in Hong Kong. There is a Technology Supporting Centre providing the local companies a variety of equipment for product development, so as to create and develop the AI ecosystem in Science Park.

The support from the government in technology and innovation is very important. Hong Kong has strong universities and technology backgrounds. The government has put lots of resources into the science development, for hundreds of billions Hong Kong dollars in the past five years. Two billion dollars will be put into the reindustrialisation fund and another two billion dollars will be put into building a Microelectronics Centre. The resources for the microelectronic industry will provide great opportunities for local companies.

MAKING MOST ECONOMIC VALUE FROM AI CHIPS DEVELOPMENT

The application of AI in various industries has been widely seen. To support the AI development, AI chip technology is the foundation and essential for continuous innovations. Hong Kong is an international city where funds, talents and information can be moving and circulating freely, providing a favourable platform and conditions for R&D and talent development, according to Dr. Mei Kei Ieong, CEO of United Microelectronics Centre (Hong Kong) Limited.

The Hong Kong government will support the local microelectronic industry development by providing a huge amount of fund in the coming years. In view of this, UMEC(HK) not only devotes much effort on R&D activities on AI chips and systems, but also promotes mutual cooperation among the academia, research institutions and industry stakeholders to support the ecosystem of innovations. It is expected that maximum economic value can be generated through close collaboration of different parties. AI-related events with a wide range of topics will be hosted by UMEC(HK) every quarter to promote knowledge and experience exchange and stimulate the industry development.

Dr. Mei Kei Ieong,
CEO of United Microelectronics Centre
(Hong Kong) Limited

Keynote 1:

New IC Design Approaches Enable AI Applications

Dr. Walden Rhines, CEO Emeritus, Mentor, a Siemens Business

Keynote 1: New IC Design Approaches Enable AI Applications

Dr. Walden Rhines,
CEO Emeritus, Mentor, a Siemens Business

The number of publications sometimes reveals the hot topic in both industry and academia. For example, back to 2005, the hot topic was cloud computing and there were 54 thousand new articles which had been published in this area. This rule also works on recent worldwide topic: artificial intelligence and machine learning. Dr. Walden Rhines, CEO Emeritus, Mentor, a Siemens Business outlined that the number of publications reaches an amazing number, 226 thousand articles in 2018.

INVESTMENT TREND IN FABLESS SEMICONDUCTOR COMPANIES HAS REVERSED

According to the statistics, venture capital investment in fabless semiconductor start-ups was on steady decline until 2017. Surprisingly, in the next year, venture capital fabless funding was setting new records by reaching $3.4 billion. Another thing worthy of mention is the portion of funding coming from China surged in recent years, exceeding U.S. in 2017.

DOMAIN SPECIFIC PROCESSORS FOR AI APPLICATIONS ARE DRIVING A NEW WAVE OF SEMICONDUCTOR GROWTH

Research confirms that traditional Von Neumann computer architectures are not efficient for pattern recognition. More specifically, computer architectures are a long way from the human brain in terms of pattern recognition and power dissipation.

As a result, nowadays, the number of neural network blocks is increasingly large in customized ASIC chips. It seems that neural networks have become a fundamental building block for AI-related machine learning.

Keynote 1: New IC Design Approaches Enable AI Applications

ARTIFICIAL INTELLIGENCE IS NOT A NEW TREND

In fact, artificial intelligence is not a new topic. There are many reasons for the delay of AI adoption in the 1980s. The most critical factor is a lack of applications to make money. Besides, there is a lack of big data for analysis and algorithms are not advanced enough to drive the AI development. In addition, the traditional computer chip architectures also limits the advance of AI. With the technology advances, human beings are in a new era with new opportunities.

MAJORITY OF VENTURE CAPITAL FUNDED STARTUPS ARE FOCUSED ON AI AND MACHINE LEARNING

With respect to the statistics on the first three rounds of venture capital funding for worldwide fabless companies, the domain specific architecture, i.e. AI and machine learning, dominates the funding for these start-ups from 2012 to 2019. In 2018, 30 AI fabless semiconductor companies were established.

The largest portion of the investment on AI is pattern recognition. For example, SenseTime is one of the world's most valuable artificial intelligence start-ups with a valuation of more than $4.5 Billion. Notice that the statistics of venture capital investment on pattern recognition have not yet covered some highly established companies like Microsoft which is now developing a custom deep neural network chip for holographic rendering.

The second largest portion of the investment is targeted at data centres for data analytics and improved processing of the collected data into useful information. For instance, nVidia, has made a remarkable transition basically introducing parallelism to the data centre processing and in fact gaining a remarkable share of data centre semiconductor revenue from it.

The third is edge computing. Today, we have the cloud, a variety of gateways and the edge nodes that provide many opportunities for start-ups and fabless semiconductor companies. However, the IoT edge nodes require technology integration, which brings a big challenge to the design industry as well as the manufacturing industry.

SYSTEMS COMPANIES ARE DOING THEIR OWN DESIGNS

The interesting part is what systems companies are doing and this is a major transition in the semiconductor world. The fact is IoT has gotten a lot of attention but semiconductor companies do not feel much of the benefit. The main reason is that the value of IoT comes from information collection and analysis. Making components is a hard way to make money. On the contrary, collecting the data, analysing it and selling it back to the people who gave it to you is a very good business. A good example to illustrate is Google. This company designed a chip to collect medical information and what they want to sell is not the chip itself but the information.

Today, the IT companies like Amazon, Facebook, Google and Alibaba have become some of the biggest customers of electronic design automation. All these companies are designing a lot of big, complicated, AI chips. Besides, 483 companies will introduce electric cars and light trucks and 255 companies have announced autonomous driving programs.

There are also some changes. One is that integrated circuits are capturing an increasing share of electronic system product value. Another is systems companies have come into the business of integrated circuit design and development. Their share has been growing at a compound growth rate of 70 percent. 17 percent of all silicon wafers produced from foundries are purchased by systems companies. They are doing their own designs, buying wafers and using those components for one thing or another.

DOMAIN SPECIFIC ARCHITECTURES REQUIRE NEW DESIGN METHODOLOGIES AND TOOLS

Several years ago, only the big, rich companies were going to be able to design with 7nm. We thought that this would be a game only for the big companies and most independent vendors would be pushed out. However, now dozens of companies with VC funding are doing very complex chips. The biggest reason is that the world is moving to the next higher level of design abstraction. This happens about every 30 years. For example, we moved from transistors to gates to schematic capture, RTL (Verilog, VHDL or System Verilog) and now finally to design with a high-level language like C++. AI and machine learning benefit from moving to the next level of abstraction.

The high-level synthesis (HLS) which has been developed by the EDA industry is an important new design tool. It separates functionality from implementation. Designers can leverage HLS to explore and deliver more with higher quality than those writing with VHDL or Verilog. HLS has many advantages. First, it improves power estimation and optimization by 30% on average. Second, HLS is easy to retarget to multiple technologies. The same C code can be retargeted for various technologies and switched between technologies easily. It can quickly explore possible alternative technologies and determine the optimal implementation. Recently, by using HLS, nVidia cut verification costs by 80%. Another example is Google which designed a VP9 CODEC with HLS in half of the normal time. The last example to show the power of HLS is CPqD which used HLS to optimize the conversion of a 28nm to

16nm design and thus saved 50% power and cost.

MACHINE LEARNING IS CHANGING EDA TOOLS AND METHODOLOGIES

EDA is rich with disruptive machine learning opportunities. For example, optical proximity correction, a technique that changes the polygon shapes so that 193nm exposure provides geometries that are about 14nm. Machine learning techniques can predict OPC output within single nm accuracy at 3X faster runtimes. Machine learning can also be exploited into other EDA applications like PCB signal integrity, simulation, process modeling and failure analysis. Besides, AI accelerators are ideal for hierarchical test. Wafer level packaging and chiplets will drive further integration.

CONCLUSION: "IT IS A GREAT TIME TO BE IN THE ELECTRONICS INDUSTRY"

The wave of the semiconductor industry is a growth wave. New technology is ready and can provide new sources of revenue for the industry. With AI, IoT, new design capabilities, and new participants in design, we are in the next wave and have the opportunity to innovate the design and architectures of new electronic products. It is a great time to be in the electronics industry.

Keynote 2:

Computing with Heterogeneous Hardware Systems for the AI Revolution

Prof. Deming Chen, Professor,
University of Illinois Urbana-Champaign

Keynote 2: Computing with Heterogeneous Hardware Systems for the AI Revolution

Prof. Deming Chen, Professor,
University of Illinois Urbana-Champaign

Heterogeneous hardware systems are becoming more and more popular in the AI era, while also facing many challenges. Prof. Deming Chen, Professor of University of Illinois Urbana-Champaign has focused on this topic for many years and attained many achievements.

About 50 years ago, AI did not take off but nowadays AI is becoming more and more popular. Many forces are now turning around and working together with AI, including big data, cloud computing, deep neural networks and hardware advances. There are lots of AI applications, ranging from autonomous driving to professional assistants. Google, Microsoft and Amazon, etc. provide us related APIs. Some hardware platforms with semiconductor backbones are also developed.

Although we have these developments, there are still some challenges, including power wall, memory wall and efficiency challenges. As the scale increases, the frequency is no longer increasing any more. The multi-level hierarchy memory improves the overall performance but the energy to bring data between registers, caches and main memory are much higher. To handle these challenges and optimize the movement of data, we can bring computations closer to the data.

With the rise of AI accelerators, the domain-specific application architectures flourish again. There are two famous examples, one is a commercial TPU from Google and another one is an academic example Eyeriss from MIT. To efficiently program the accelerators and integrate them into the common systems, reconfigurable computers are proposed as a solution.

Keynote 2: Computing with Heterogeneous Hardware Systems for the AI Revolution

RECONFIGURABLE COMPUTING

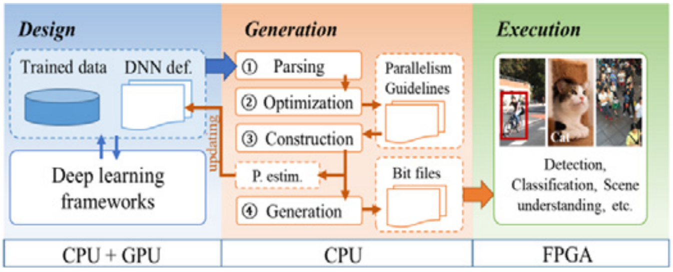
Reconfigurable computing is a computing architecture that tries to blend the flexibility of software with the performance and efficiency of custom hardware. Although not as efficient as ASICs, it is very flexible. Many evolving approaches are therefore made to improve the efficiency.

For example, Xilinx proposes the Versal ACAP architecture, with some DSPs on it and embedded AI engines. To some extent, this kind of architecture is no longer Field Programmable Gate Array, it can be called Field Programmable ASIC Array. This architecture converges the flexibility and efficiency. There are also some other works from Intel and Wave Computing. DARPA is also pushing several hardware projects.

In industry landscape of DNN accelerators. CPU and GPU are more flexible. ASICs are more efficient. Soft DPU is in the middle. FPGA is one example of soft DPU. FPGA is not only for computation, but also for communication. It offers the high performance and flexibility. We can deal with changing algorithms because of its high reconfigurability.

DNNBuilder – Building DNN from Ground UP

DNNBuilder is an open-source framework, in collaboration with IBM. For DNN applications, the basic building blocks are regular. Borrowing the idea from the design of SoC, we can reuse some IPs. A bunch of IPs are developed, for convolutional operations, fully connected layers and some other operations. Based on different deep neural networks, we can quickly analyse what kind of IPs are needed. Then we can quickly map these models to FPGA. After getting the feedback from the FPGA, the architecture is automatically rebuilt according to the performance and resource estimation. If this design is not ideal, we can go back to the algorithm side and optimize the algorithm design.



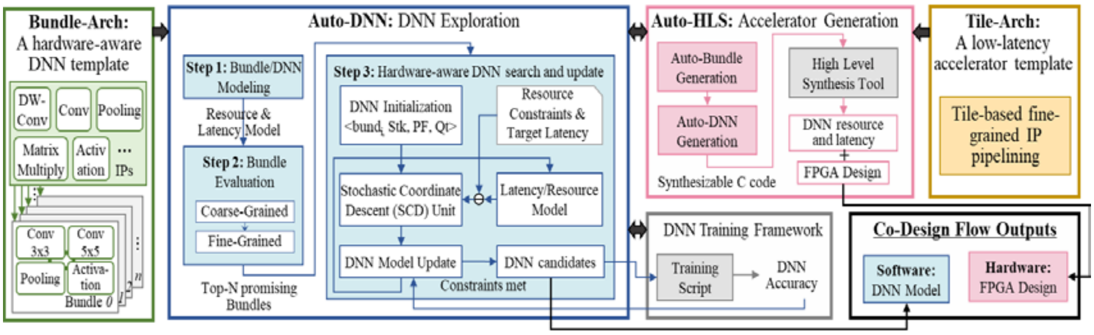
In DNNBuilder, the kernel, channel and bit-width are reconfigurable. The Cache architecture is also used in DNNBuilder. Four slices cache on-chip are kept on board instead of keeping the whole feature maps. This kind of design can reduce 7.7x latency for running YOLO. DNNBuilder can also save the usages of BRAM by more 100x.

Cloud-DNN

Cloud-DNN is an open framework for mapping DNN models to cloud FPGAs. Given a Deep neural network, we analyse each layer and then translate that layer in C-level functions. Then we call HLS tools to generate RTL and then map the whole model to the cloud FPGA.

Keynote 2: Computing with Heterogeneous Hardware Systems for the AI Revolution

FPGA/DNN Co-design Framework for IoT Intelligence



FPGA/DNN co-design framework is for IoT intelligence. Mostly when scientists design their deep neural networks, they pay no attention to the future platforms. For edge devices, GPU, TPU and FPGA, the hardware constraints are totally different. In this framework we want to design the DNN model and hardware together. In other words, this is the hardware and software co-design. We have a bunch of hardware-friendly building blocks. We have engines which can generate RTL automatically and produce the co-design solution.

T-DLA: An Open-source FPGA Accelerator for Ternarized DNN Models

We also developed an open-source FPGA accelerator for Ternarized DNN models. Binary neural network is a lightweight model. The drawback is that the accuracy loss is high. In the ternarized model, lots of designs are customized with high flexibilities.

The goal of future computing is to improve the efficiency under the same transistor count. Reconfigurable accelerator design is a new solution and promising. For heterogeneous computing, new compiler techniques for mapping DNNs onto FPGAs to meet the high demand for energy efficiency and performance.

CONCLUSION

The goal of the future computing is to improve efficiency under the same transistor count. Reconfigurable accelerators and heterogeneous computing are promising. We can promote the development of AI accelerators from different levels, including chip level, system level, edge and cloud level. Current compiler techniques are not powerful enough to map DNNs onto FPGAs therefore new techniques are needed to meet the high demand for energy efficiency and performance. Now it is an exciting time for realizing the true potential of AI and novel architecture design.

Keynote 3:

AI in Physical Design Tools

Prof. Evangeline Young, Professor,
Department of Computer Science and Engineering,
The Chinese University of Hong Kong

Keynote 3: AI in Physical Design Tools

Prof. Evangeline Young, Professor, Department of Computer Science and Engineering, The Chinese University of Hong Kong

Recently, AI techniques are heavily introduced in the whole IC design flow including physical design stage. They offer new solutions to address conventional problems in a more effective and efficient way. Prof. Evangeline Young, Professor of the Department of Computer Science and Engineering at The Chinese University of Hong Kong is a leading expert on physical synthesis, outlining the recent notable progress by using AI techniques in physical design.

THE OVERVIEW OF PHYSICAL DESIGN UNDER CURRENT TECHNOLOGY NODE

The IC design flow includes many stages starting from system specification to packaging and testing. In the physical design stage, many steps such as floorplanning, placement, clock tree synthesis, signal routing, timing and DFM closure are involved. These steps have been studied for many years. On the other hand, we have all witnessed the continuous shrinkage of the size of semiconductors over past decades. This aggressive scaling brings many big challenges to existing physical design algorithms. AI techniques may be a cure.

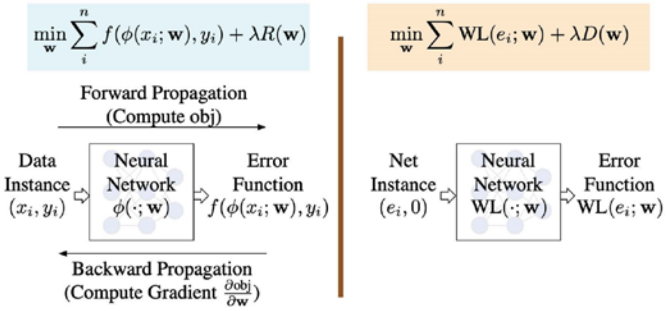
A PLACEMENT EXAMPLE

A recent academic placer called "DREAMPlace" casts the placement problem to a deep learning problem. With advancement in AI hardware and software, it makes a good example to illustrate how AI can help with physical design.

Placement is a critical yet time-consuming step in physical design. Since it determines the locations of standard cells and macros in the physical layout, its quality has huge impacts on the consequent steps in the flow such as routing and post-layout optimization. The commercial design flows often run core placement engines many times to achieve design closure. As placement involves large-scale numerical optimization, conventional analytical placer, which essentially is solving complex nonlinear optimization problems, usually takes hours for large designs, which will slow down the design iterations.

"DREAMPlace" is a a GPU-accelerated analytical placer obtained by casting the analytical placement problem to training a neural network. The analogy between neural network (NN) training and placement is shown as follows.

Keynote 3: AI in Physical Design Tools


$\min_{\mathbf{w}} \sum_{i}^{n} f(\phi(x_i;\mathbf{w}),y_i)+\lambda R(\mathbf{w})$
Forward Propagation
(Compute obj)
Data Instance (x_i, y_i)
Neural Network $\phi(\cdot;\mathbf{w})$
Error Function $f(\phi(x_i;\mathbf{w}),y_i)$
Backward Propagation
(Compute Gradient $\frac{\partial \text{obj}}{\partial \mathbf{w}}$)
$\min_{\mathbf{w}} \sum_{i}^{n} \mathrm{WL}(e_i;\mathbf{w})+\lambda D(\mathbf{w})$
Net Instance $(e_i, 0)$
Neural Network $\mathrm{WL}(\cdot;\mathbf{w})$
Error Function $\mathrm{WL}(e_i;\mathbf{w})$

(a) Train a NN

(b) Solve a placement

A ROUTING EXAMPLE

Congestion estimation plays an important role in global routing (GR), which benefits the detailed routing step. The study on using deep learning technology to predict GR congestion opens a new direction to avoid DRC violations at an early stage.

With features collected before GR like standard cell density map and target label as a 2-D binary map reflecting GR congestion, a fully convolutional network (FCN) is exploited as the prediction model.

Experimental results demonstrate that the deep learning-based framework successfully improves the QoR of GR solutions, leading to less design rule violations after detailed routing.

A MASK SYNTHESIS EXAMPLE

OPC is one of the representative resolution enhancement techniques (RETs) that can ensure mask printability thus improve chip manufacturing yield. However, in advanced technology nodes, the conventional mask optimization process called inverse lithography technology (ILT), consumes increasingly more computational resources. A new end-to-end mask optimization framework that combines UNet and ILT is proposed to achieve a much faster convergence rate with improved quality.

CONCLUSION: "AI BRINGS OPPORTUNITIES AND CHALLENGES TO PHYSICAL DESIGN. INTELLIGENT AND FAST PHYSICAL DESIGN TOOLS HELP DESIGN OF AI CHIPS"

AI techniques, especially deep learning, offer opportunities for recasting conventional physical design problems in new directions so that more efficient design tools can be achieved. The better EDA design tools will benefit the design process of AI chips. The AI chips finally advance the development of AI techniques, which power the design tools.

Keynote 4:

Machine Learning for Chip, Package and Board Design

Prof. Elyse Rosenbaum, Professor,
University of Illinois Urbana-Champaign

Keynote 4: Machine Learning for Chip, Package and Board Design

Nowadays, EDA tools are amazing. They allow people to successfully design and manufacture chips with over a billion of transistors on them. Nevertheless, they are not absolutely perfect, according to Prof. Elyse Rosenbaum, Professor of University of Illinois Urbana-Champaign, who works in CAEML. CAEML is a NSF Industry/University Cooperative Research Centre of three universities, University of Illinois at Urbana-Champaign, Georgia Tech and North Carolina State University. It enables fast, accurate design and verification of microelectronic circuits and systems by creating machine-learning algorithms to derive models used for electronic design automation.

THE LIMITATION OF CURRENT EDA TOOLS

EDA tools are not perfect, because not all chip designs work on the first silicon spin and chips failure on qualification testing can usually be traced back to an insufficient modelling capability. Furthermore, they cannot be used for design optimization because the number of design variables is so huge that people cannot fully explore the design space using simulation. As a result, people design the chips of their products to meet specifications rather than executing the absolute optimal design.

1] Harry Foster, 2012 Wilson Research
http://www.mentor.com/products/fv/m
ER PROPRIETARY
ght-leadership Series
Summit
AI

Prof. Elyse Rosenbaum, Professor, University of Illinois Urbana-Champaign

WHY SHOULD WE USE MACHINE LEARNING?

Circuit designers, semiconductor device specialists, and some signal integrity engineers often need to find the minimum of some non-convex functions when they are fitting a model or optimizing the design. Furthermore, they often need to construct stochastic models because of manufacturing variations. Machine learning specialists perform similar numerical analysis and the code is available. Turns out that doing design optimization is even more challenging than optimizing or fitting a model because of dimensionality, where the number of design variables is very large. When doing design optimization, they usually do not know the function that maps from the feature space or the input variable space to the final performance of the design, that means they do not have gradient Information available. Therefore, they cannot use algorithms such as the cast of gradient descent.

Electrical engineers usually work with what the statisticians called discriminative models. What this discriminative model does is it gives the expected value of the output for a given input. A generative model, in contrast, gives the joint probability distribution between the inputs and the outputs. Engineers can sample from that distribution, thereby generating synthetic new samples. If we want to do yield analysis in light of manufacturing variations, we need those generative models.

What's more, it is very important that we can characterize this statistical distribution for our yield analysis and we are going to be dealing with a very large dimensionality data space. As the simplest non-parameters way to characterize the resulting statistical distribution, the methods like kernel density estimation will be inefficient due to the size of model growing with the amount of training data. Therefore, it will be much better if we can use a parametric model such as a neural network.

GAN

Absolutely, state of the art in neural network based generative models is the GAN, the generative adversarial network. Take the human face generator GAN as an example. We have a neural network called the generator, which maps from a sample in this native space to the space of human face images. We have another neural network that we are training simultaneously called the discriminator. What it does is to compare the images created by our generator neural network to images in a real database of photos and determines whether it is a fake image or a real image. Once the discriminator can no longer tell if an image is real or fake, we finish training our model.

BEHAVIORAL MODEL

This kind of model is used to verify a system design prior to manufacturing.

Behavioural model is only going to represent the response of your component at its external ports, where it connects to other components of your system. It is typically less complex than models that reveal the inner workings of the component and gives great savings on computational efficiency. Furthermore, behavioural models by obscuring the inner workings of the component protect the designer's intellectual property. Also, it is very important to validate the machine learning behavioural model and make sure all of its predictions obey the laws of physics in order to model real physical hardware.

Recurrent neural network is a very useful class of behavioural models for representing individual IP blocks. The IP block was in the form of an encrypted net list. People have no idea what the circuit was, but they know it contained over a thousand MOS transistors, each of which was represented by a piece of full model. CAEML researchers trained the RNN and they did transistor level simulations, RNN simulations. The result shows that the discrepancy between the two sets of predictions was less than 1% and the RNN model simulated more than 40 times faster than the transistor level netlist and time savings grow as the size of the netlist increases.

APPLYING MACHINE LEARNING TO EDA

1. Thermal design optimization for 3D IC

Self-heating effects tend to be quite severe in 3D IC and the thermal gradient is larger in Die Stacking. A large thermal gradient directly translates to an increased clock skew and the package physical design has a strong influence on those thermal gradients. Design Space Exploration is done using very slow computational complex physical simulations. So instead, using Bayesian optimization techniques to find where in this design space can get the smallest skew.

The CAEML researcher at Georgia Tech developed a new optimization technique called Two-Stage Bayesian Optimization (TSBO). One of its most important features is doing online active learning of the acquisition function, which can help people to work with the acquisition function that is most ideally suited to the problem at hand. Secondly, instead of optimizing the acquisition function, TSBO simply evaluated at a small number of points. According to the experimental result, this optimum method converges towards the optimum solution much more quickly even though it is not the fastest algorithm compared with other state of the art optimization algorithms. Consequently, the number of physical simulations is minimized.

2. Placement-to-Clock-Tree prediction

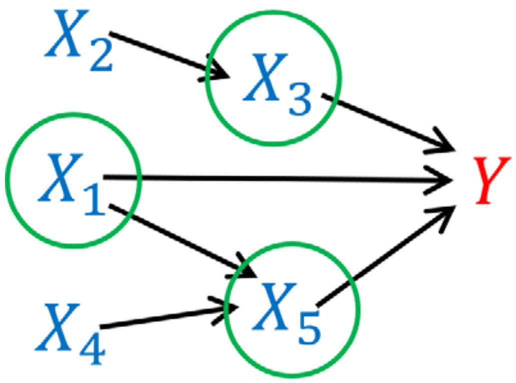
For one candidate placement which predicts without actually synthesizing and simulating the clock tree and predicts what will be its quality in terms of wirelength, skew and power. CAEML uses GAN for prediction, that inputs to the model are images taken from a large database.

Firstly, it includes a regression model that takes a picture of whose inputs are the netlist, the candidate flip-flop placement, and clock tree synthesis parameters and it predicts the resulting power, wirelength and skew. The other part is CAEML has a generator which gives inputs of the flip-flop placement and the netlist, will output examples of good clock tree synthesis setting. When they use this GAN clock tree synthesizer, the design it suggests is better than that obtained using the commercial EDA tools. And for a previously unseen netlist that was not part of the training data set, it can achieve 88% fewer buffers, 5% less skew and 52% less power.

3. Data analytics

Prof. Rosenbaum showed how to use machine learning models for prediction of the impending failure of a Hard-Disk Drive or a Solid-State Drive. These SSD and the HDD already have built in sensors and this technology is called SMART, self-monitoring analysis and reporting technology. The objective is to maximize the failure detection rate, subject to the requirement that the false alarm rate cannot exceed 1% and to predict the failure at least one day in advance of what it actually occurs, so that the operator can take some corrective action.

CAEML uses a GRU model, gated recurrent unit model to finish the data analytics and performs feature selection to get the minimal set of prognostic model inputs because there are too many available SMART attributes. CAEML employs a causal feature selection approach. Y is the output of our model and the Xes are the inputs. Directive Information Graph is shown as follows.



The Xi is not a direct cause of Y if Y becomes independent of the Xi once it is conditioned on all the other features. CAEML just selects these direct causes as model inputs. Experiments have shown that this is effective for decreasing feature numbers and getting a statistically indistinguishable result.

CONCLUSION

Using machine learning algorithms can help overcome the limitations of modern EDA tools. CAEML researchers applied machine learning to EDA in 3 aspects, thermal design optimization, Placement-to-Clock-Tree prediction and data analytics, successfully improving the efficiency and achieving more optimised result.

Panel Discussion:

New EDA tools? Beyond traditional EDA, what is the chance for start-ups?

Moderator:

Prof. Philip Chan, Deputy President and Provost,
The Hong Kong Polytechnic University

Panellists:

Dr. Walden Rhines, CEO Emeritus, Mentor, a Siemens Business

Prof. Deming Chen, Professor,
University of Illinois Urbana-Champaign

Prof. Martin Wong, Dean of Engineering,
The Chinese University of Hong Kong

Prof. Frank He, Director,
The Key Laboratory of Integrated Microsystems, Peking university

Dr. Mei Kei Ieong, CEO,
United Microelectronics Centre (Hong Kong) Limited

1. In the context of AI chip design, do we need new EDA tools, e.g. High Level synthesis tools?

Prof. Frank He, Director of The Key Laboratory of Integrated Microsystems at Peking University concurred that the traditional EDA tools need to improve themselves by making some fundamental changes based on AI methodologies.

Dr. Walden Rhines, CEO Emeritus, Mentor, a Siemens Business said that the demand of new AI tools makes the business great. He observed that according to the performance metrics, the biggest bottleneck in specific domain of AI is memory access which contributes the most to the latency.

2. Von Neumann Architecture is not efficient for AI applications. Do we need new architecture level tools, e.g. planning tools and memory hierarchy optimization tools?

Prof. Martin Wong, Dean of Engineering at The Chinese University of Hong Kong thought it is necessary for industry and academia to develop new architecture tools. He took the deep neural network as an example to illustrate his point. Without the understanding of real structures of networks, existing EDA tools cannot give a feasible solution to placement, routing, and floorplanning tasks. With prior from layers and memories, these tools can be enhanced. With more understanding of the underlying domain, we will be able to come up with a better design.

Prof. He viewed this problem from domain shift and mapping AI application from Python to FPGA. Currently hardware researchers focus on High Level Synthesis and C/C++ library OpenCL. For data scientists, they adopt TensorFlow, Pytorch and Caffe as their developing frameworks. However, the aforementioned frameworks are based on Python. Therefore, mapping them to low level HLS for FPGA is urgent.

Dr. Mei Kei Ieong, CEO of UMEC(HK) expressed that AI techniques bring more opportunities for the small companies and start-ups to develop new design tools. Previous design software are offered by top three companies which dominate the whole market. With introducing AI techniques into the EDA domain, small companies and start-ups with large venture capital investments can survive and have promising future.

3. Heterogeneous system integration is the new trend for AI chips. Do you think we can use existing functions and packing technology to implement the whole heterogeneous system?

According to Dr. Rhines, the traditional race to monolithic compression to a single chip does not make sense if the cost per unit area is dramatically greater for the next technology node. Reusing functions that can be implemented in older nodes is potentially more efficient.

4. Do the current EDA tools support heterogeneous integration?

Dr. Rhines said that in the last decade there has been large progress on heterogeneous integration, including design planning, thermal analysis, performance prediction etc. Nevertheless, the design community still has a long way to go. He takes the thermal analysis as an example. Under dynamic conditions, the thermal conditions of heterogeneous voltage package are far from uniform. Thermal activity is concentrated in small areas that affects the electrical performance as well as viability.

Prof. Deming, Professor of University of Illinois Urbana-Champaign gave his point from the perspective of High Level Synthesis. Different applications in EDA domain have different demands on heat and performance. HLS can handle various hardware descriptions easily to find the optimized designs, which can rewrite desired RTL. Meanwhile, HLS can also tackle different architectures.

Dr. Ieong supplemented that heterogeneous integration can give small companies and start-ups more opportunities to take part in the design of new EDA tools.

5. With the increased use of AI techniques and technologies, is there any needs to reduce the roles of design engineers?

Prof. Wong claimed that there is no need to reduce the role the engineers. In the design automation projects conducted by DARPA, no engineers take part in the design loop. However, there are not a lot of people losing their jobs. Because you still need lots of people in the loop to write the programs. Another point given by Prof. Wong to support his point was that with aggressive scaling of size of semiconductors, the development of EDA tools is always behind the advance of technology nodes. More and more people are expected to be involved in the EDA industry.

Prof. Philip Chan, Deputy President and Provost of The Hong Kong Polytechnic University, agreed with Prof. Wong's opinion. He pointed out although HLS can help reduce lots of engineering work, more emerging problems need more efforts from human experts. Prof. Chan gave us mask optimization under advanced technology nodes as an example.

Dr. Rhines also supported Prof. Wong's view. We continue to raise the complexity and the number of design rules at a much faster rate than we can improve the methodology and tools. He thought that there is no danger of this happening.

6. What are the opportunities for starting AI chip companies? Which area can a start-up company do well in?

Dr. Rhines thought that things are different from the past. Because more and more AI companies becomes the potential clients for AI chip companies including fabless semiconductor start-ups. There is also a large opportunity to improve the design tools and methodologies to satisfy the special requests from AI. Perhaps special silicon is also required.

Dr. Ieong thinks that successful start-ups should have several important ingredients. First, traditional architectures of products should be changed. Second, aggregation is needed. Many companies get used to doing their own things now, which leads to their failures in markets. Third, successful companies should not be isolated and still rely on traditional partnership model.

7. What are your opinions on open-source EDA tools?

Prof. Chan said that there is a trade-off between commercial tools and open-source tools. For companies who want to have reliable supports, maintenance and bug fixes, they prefer developing their own tools or using commercial tools. Sometimes open-source tools can also satisfy our requirements.

From the perspective of Dr. Rhines, open source tools require a huge amount of human resources and investment to become usable and reliable. RISC-V and RedHat are two representative examples.

Dr. Ieong noticed that the companies may not buy unreliable or newly developed tools. To promote the developments of these EDA tools, he urged the government to fund universities and start-ups which are upgrading the EDA tools. That will be a golden era for developing open source EDA tools.

8. Is there any new opportunities for downstream manufacturing companies and packing companies? What is the impact of AI on the downstream new players?

Dr. Ieong claimed that more investment from the government will promote the companies from the downstream and help the whole community to lower the technology barrier.

Prof. Chan stated that the heterogeneous integration will be a powerful strategy for small companies and these new players. Deep learning industry is a representative integration which combines many things like sensors, IoT, etc. This is a potential new playground for heterogeneous integration.

Keynote 5:

Efficient Reconfigurable Hardware for AI

Dr. Hayden So, Associate Professor,
Department of Electrical and Electronic Engineering,
The University of Hong Kong

Keynote 5: Efficient Reconfigurable Hardware for AI

Dr. Hayden So, Associate Professor of Department of Electrical and Electronic Engineering at The University of Hong Kong presented the keynote speech on reconfigurable hardware for AI.

To begin with, reconfigurable technology is one of the key technologies that will enable AI applications in the coming years. It is recognized as a viable way to accelerate the AI applications, according to Dr. Hayden So, who has been in reconfigurable technologies area for more than 15 years. Dr. Hayden So shared the experiences of implementing accelerators efficiently with the help of using FPGA overlays.

FPGA OVERLAY

The idea of FPGA overlay is to design a virtual architecture that can serve as a bridge between the applications and the physical hardware. Overlays can be used for many purposes, such as to improve design portability, improve designer productivity, and to assist with hardware debugging. As a means to improve design productivity, the additional overlay facilitates separation of concerns between the high-level algorithm designers and the low-level physical implementation engineers. With the overlay as a rapid compilation target, high-level algorithm designers can focus solely on identifying performance bottlenecks for acceleration, while leaving the low-level optimization of the overlay to the specialized hardware design team. Overlay can thus help to accelerate the process of designing an accelerator by improving the "Design-Implement-Debug" cycle, and facilitates design-space exploration for the target applications. Finally, the optimized design may serve as the basis for application-specific integrated circuit (ASIC) implementation.

Dr. Hayden So, Associate Professor, Department of Electrical and Electronic Engineering, The University of Hong Kong

Two examples of FPGA overlay were presented by Dr. So.

1. QuickDough

The main idea of QuickDough is to facilitate the design of hardware/software systems for accelerating tight loops in high-level source codes by using a soft coarse-grained reconfigurable array (CGRA) overlay. With this framework, high-level software is compiled onto the overlay rather than the physical FPGAs. As a result, the conventional design flow of hardware, software, and hardware/software co-design are separated and transformed into a unified software compilation framework. A typical design process in QuickDough consists of three stages. In the first stage, designers utilize the fast compilation path from the whole design flow, which is essentially a software scheduling problem, to make rapid prototype of the target hardware/software system. In the second stage, the designers may further improve system performance by devoting additional design time on exploring different overlay architectures. In case a new overlay architecture is desired, designers may proceed with the third stage that involves the physical design of a new overlay. The whole process benefits from the fact that the FPGA overlay is a virtual architecture, which does not need to be built in hardware and thus ensures the flexibility to be changed anytime according to the requirement. This is an excellent explanation of FPGA overlay as a solution that allows rapid product design cycles to changing requirements.

In short, the use of overlay is a different way of rethinking the whole hardware-software design flow problem. It allows fast architecture exploration that is essential for good overall system design. High-level designers benefit from the software-like design flow that takes mere seconds instead of following the conventional FPGA design flow that would take hours to days or more. Finally, designers could choose different levels of optimization by adjusting the reconfigurable array.

2. GraVF: Distributed Graph Processing on FPGAs

Besides neural network and deep learning, graph algorithm is another important category of algorithms that can benefit from hardware acceleration. GraVF is a python-based hardware generator that allows user to create large-scale graph processing systems that run on clusters on FPGAs. The powerful thing about GraVF is its ability to generate very large-scale hardware with a simple high-level vertex-centric graph processing model. Using the vertex-centric processing model, users only need to provide GraVF with the content, timing, and distribution rule of the messages for each vertex. The framework will subsequently produce all the necessary hardware infrastructure for vertex partitioning, message transmission and synchronization automatically.

FUTURE RESEARCH

Before concluding, Dr. So also touched on 2 additional on-going projects regarding AI acceleration on FPGAs. The first was NITI, which was a novel integer-only neural network training framework. Unlike typical neural network training frameworks that rely on floating-point arithmetic, NITI uses only integers throughout the algorithm, making it suitable for energy-efficient hardware implementation. The second was FTDL, which was a family of scalable overlays designed for efficient deep neural network inference on FPGAs.

CONCLUSION

Overlay can effectively facilitate many aspects of the design process, including the initial design, implementation, evaluation and debug. The challenges for future research are to design very efficient domain-specific overlay architectures that can map onto different, including emerging, physical hardware platforms.

Keynote 6:

Energy efficient AI accelerators for sensor-based edge devices

Prof. Chi Ying Tsui, Professor,
Department of Electrical and Electronic Engineering,
The Hong Kong University of Science and Technology

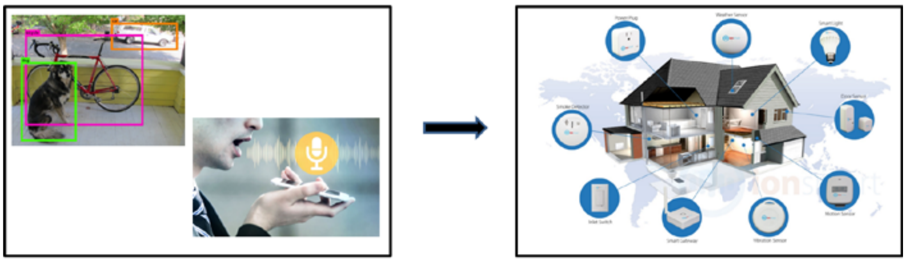
Keynote 6: Energy efficient AI accelerators for sensor-based edge devices

This keynote was presented by Prof. Chi Ying Tsui, Professor of Department of Electrical and Electronic Engineering at The Hong Kong University of Science and Technology. In this talk, Prof. Tsui proposed methods for reducing the power consumption in running convolutional neural networks (CNNs) such that advanced AI techniques can be deployed on edge devices instead of on the cloud. Through exploiting edge AI computing, significant energy and memory bandwidth savings can be achieved, while promoting a low latency due to localized computing. Three ways were highlighted by Prof. Tsui to achieve these goals.

SPARSITY-BASED ENERGY-EFFICIENT ARCHITECTURE

One way to reduce the computational complexity in CNNs is by exploiting the weight and activation sparsity of the neural networks. As is well-known, the node operation in a neural network mainly consists of weighted summation of connected neuron inputs and the nonlinear activation (typically ReLU). Due to the nature of ReLU, negative input will produce a zero output, thus computing that negative input is redundant. Even more, such redundancy accounts for approximately 50% in the overall network inference computation. To address this, Prof. Tsui proposed LRADNN comprising two phases. First, LRADNN predicts the activeness (i.e., zero or nonzero) of each neuron to be computed, called the prediction phase. Second, in the feedforward phase, only the predicted-nonzero neuron inputs are actually calculated. Building upon this idea, Prof. Tsui proposed two additional ways to reduce the computation. One way is by adding the low-rank constraint in a weight matrix. Another way is by designing a distributed NoC-based architecture

Prof. Chi Ying Tsui, Professor,
Department of Electrical and Electronic Engineering,
The Hong Kong University of Science and Technology



- Almost all recognition and detection tasks
- Cloud-based solution

- Save BW, power, latency
- Running **inference** locally

Figure K6-1. Demands for Low Power CNN Processor

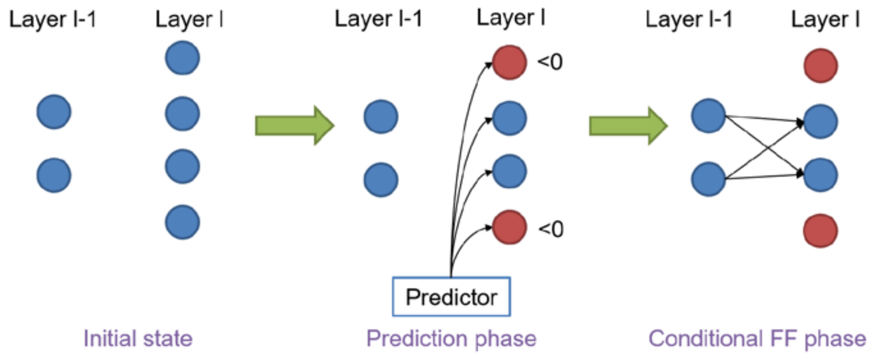


Figure K6-2. Two phases in LRADNN

to optimize the computation and memory access, which achieves a 10%~70% throughput improvement and around 50% power consumption reduction.

PRUNING AND COMPRESSION

Weight pruning is a popular way to reduce CNN computation. It is observed that a lot of the synaptic weights in deep neural networks (DNNs) have very small magnitude, and therefore these unimportant weights can be removed without sacrificing much accuracy. However, pruning is commonly done in an unstructured manner, which makes efficient implementation of the CNN inference on systolic arrays difficult. In this regard, structured pruning is desired, which aims at pruning a whole row or column of the weight matrix. Nonetheless, deleting a whole row or column is accompanied by a significant accuracy drop. Subsequently, Prof. Tsui proposed a scheme to combine several sparse columns to form a group which is then mapped to a single column of the systolic array. By doing so, the weight matrix can be substantially compressed while maintaining a high accuracy. Another way to achieve structured pruning is by using row and column permutations. Promising results have been demonstrated for these versatile schemes.

IN-MEMORY COMPUTATION: CIRCUITS AND ARCHITECTURES

Processing near or in memory is an effective way to reduce data access latency. Binarized neural networks (BNNs) constitute a good choice to facilitate in-memory computing since the arithmetic operations are replaced by one-bit operations, thus accelerating the overall computation a lot. Based on BNN, Prof. Tsui mentioned three specific works to save energy on the edge. The first one is the charge-based BNN accelerators which employ binary convolution to achieve 6X improvement in energy for CIFAR-10 dataset. Second, researchers had utilized conductance to represent the weight parameters in a BNN and designed the corresponding analog circuit to replace the matrix-vector multiplication for inference acceleration. Third, current-based SRAM

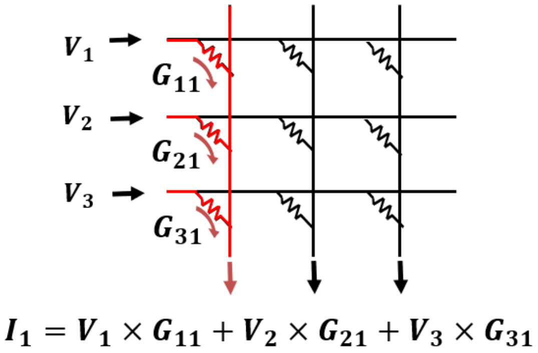

V_1
V_2
V_3
G_{11}
G_{21}
G_{31}
$I_1 = V_1 \times G_{11} + V_2 \times G_{21} + V_3 \times G_{31}$

Figure K6-3. Replacing matrix-vector multiplication with analog computation

array was employed for MAC (multiply-accumulate) operation which consumed 113X lower energy than discrete SRAM/ digital MAC for a 10-way classifier.

RESISTIVE RAM (RERAM)-BASED CNN ARCHITECTURE

The last point Prof. Tsui mentioned was to make use of resistive RAM (ReRAM) to realize the computation in CNNs. This could result in various advantages such as low voltage operation, in-memory computation and non-volatile data. ReRAM-based CNNs were demonstrated to achieve a 15X higher throughput and 5.5X lower energy than state-of-the-art conventional processing element (PE) array-based accelerators. Some other works further improved the performance by utilizing powerful techniques such as subword-based encoding, microarchitecture optimization and dynamic quantization, thus leading to an additional 2X~4X improvement in terms of throughput, energy efficiency and area efficiency.

CONCLUSION

The improvement of the AI accelerator energy efficiency needs careful consideration from all aspects. More memory- and computation-efficient networks need to be developed, that approximate computing taking into account the network characteristics is a viable solution. Besides, architecture-algorithm co-design should be enforced to fully exploit and optimize the algorithms to the hardware architecture. In-memory computing is a strong candidate for alleviating energy-hungry data movements. We should also explore emerging devices and new computing paradigms to drastically improve the energy efficiency.

Keynote 7:

The Challenge and Opportunity of Heterogeneous System Integration

Mr. Nelson Fan, Vice President,
APT Business Development, ASM Pacific Technology

Keynote 7: The Challenge and Opportunity of Heterogeneous System Integration

Mr. Nelson Fan, Vice President, APT Business Development of ASM Pacific Technology, a world leader in semiconductor assembly and packaging, gave a talk on the challenges and opportunities of Heterogeneous Integration (HI). During the presentation, Mr. Fan first introduced different approaches of 2.5D integrated circuit (IC) HI, followed by the fabrication process and technical challenges in this domain. Then, the emerging 3D IC HI technology was briefly illustrated alongside its challenges.

HETEROGENEOUS INTEGRATION – TO BE MORE THAN MOORE

Heterogeneous Integration (HI), the process of assembling and packaging multiple components on a single chip, is a critical packaging technique to boost the functionality and operating performance of ICs. With HI, components with different functionalities (e.g., processors, RF, memory modules etc.), different process technologies, and even separate manufacturers can be integrated for elevated efficiency of chips and "More than Moore" performance. As the chips scale denser and smaller, the advanced packaging technology is entering the "litho" scale (viz. nanometer) from the usually adopted millimeter and micrometer scales. As stated by Mr. Fan, packaging could be divided into three levels nowadays, namely, package level, chip level and transistor level. This talk was focused on the package level packaging.

Mr. Nelson Fan, Vice President, APT Business Development, ASM Pacific Technology

WHERE WE ARE: 2.5D HETEROGENEOUS INTEGRATION FOR HIGH-PERFORMANCE CHIPS

For 2.5D HI, there are mainly three different approaches of integration at the package level, TSV (Through-Silicon Vias) interposer, embedded Silicon bridge and HIFO (Heterogeneous Integrated Fan Out). TSV interposer is a well-known technology pioneered by TSMC, integrating different components (e.g., dice of GPU, CPU, FPGA or high-bandwidth memory) on top of a Silicon interposer. The components on the interposer can communicate with each other through fine lines and space RDL (Re-Distribution Layer) or with lower layers by TSVs. However, the integrated components, named "Compound Die" is getting bigger and bigger in size, according to Mr. Fan, towards 70x70mm2, which is a challenge to the manufacturing cost and yield due to issues such as large die size, large warpage, stress and die crack. Moreover, considering the common wafer of 300mm diameter size for the interposer, the number

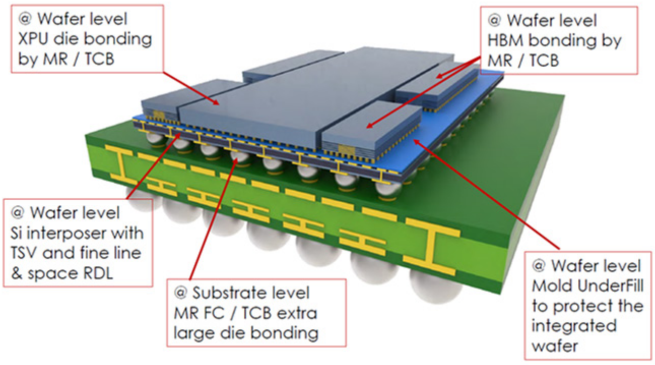

@ Wafer level XPU die bonding by MR / TCB
@ Wafer level HBM bonding by MR / TCB
@ Wafer level Si interposer with TSV and fine line & space RDL
@ Substrate level MR FC / TCB extra large die bonding
@ Wafer level Mold UnderFill to protect the integrated wafer

Figure K7-1. Illustration of a TSV

of compound dice that can be made on each wafer is quite limited by the TSV approach.

On the other hand, the embedded Silicon bridge integration utilizes high density interconnect substrate with a cavity for a high precision bridge, through which the components can communicate with each other. Mr. Fan stressed that the embedded Silicon bridge technology could save more space by omitting the second interposer distributing signals to the whole compound die, in contrast to using TSVs. Nonetheless, challenges remained for this approach, such as the coplanarity control of the bridges, ECD (Electrochemical Deposition) with different bump densities, TCB (Thermal Compression Bonding) for chips with different bump densities and placement precision etc.

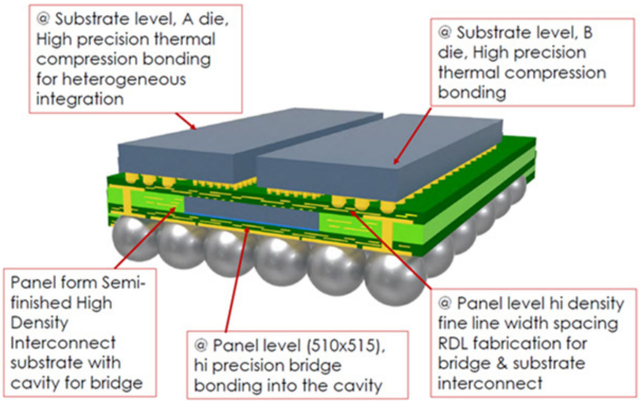

@ Substrate level, A die, High precision thermal compression bonding for heterogeneous integration
@ Substrate level, B die, High precision thermal compression bonding
Panel form Semi-finished High Density Interconnect substrate with cavity for bridge
@ Panel level (510x515), hi precision bridge bonding into the cavity
@ Panel level hi density fine line width spacing RDL fabrication for bridge & substrate interconnect

Figure K7-2. Embedded Si bridge integration

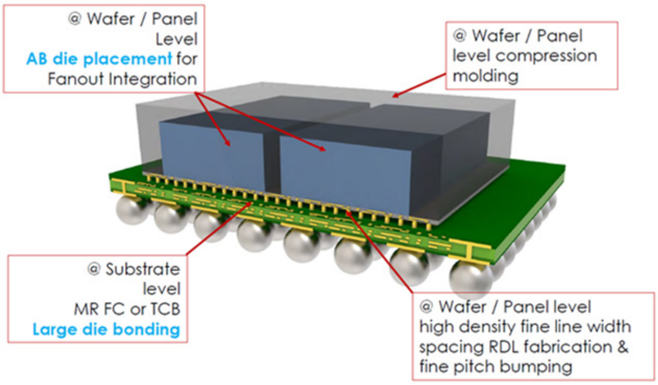

@ Wafer / Panel Level
AB die placement for Fanout Integration
@ Wafer / Panel level compression molding
@ Substrate level
MR FC or TCB
Large die bonding
@ Wafer / Panel level high density fine line width spacing RDL fabrication & fine pitch bumping

Figure K7-3. HIFO illustration

HIFO, another critical technology in HI, attempts to integrate different dice directly onto an RDL fabrication layer that is connected to the substrate via technologies such as TCB. For these three approaches of 2.5D HI, ASM Pacific Technology has developed many different types of products for IC packaging solutions, with critical technologies such as Low-k die/Si bridge for laser dicing, TCB for chip level and compound die interconnection, TSV and RDL fabrications etc. Their solutions and products are also leveraged for industrial chip production for giant companies such as Samsung, Huawei, HiSilicon etc.

TOWARDS THE FUTURE 3D HI –THE MORE FLEXIBLE SOLUTION

In the last part, Mr. Fan further mentioned the latest trend of 3D HI technology. Starting with a wafer-level active device, fabricated with TSV and CMP (Chemical Mechanical Planarization) top surface, dice from different process technologies could be placed onto it with extra-high-precision die placement tools. Then, after adding wafer level mold as well the RDL layer, all components formed an integrated compound die that could be further integrated with other devices (e.g., HBM) via different approaches such as the three above-mentioned 2.5D HI technologies. Consequently, it was believed 3D HI could enable more adaptive solutions and compatibility for components from different processes and cost consideration. In short, 3D HI technology was entering a new era that challenged the engineers of flatness, cleanness, as well as front-end & back-end technologies. The research on 3D HI was still on-going, and Mr. Fan told the audience that he was looking forward to collaboration for pushing the development of this frontier.

CONCLUSION

In summary, through Mr. Fan’s presentation, the advanced packaging technology that enables “More than Moore” has been introduced, with a brief illustration of 2.5D and 3D HI technologies which are becoming increasingly popular in real-world production. As the scale of interconnected pitches gets smaller and smaller to the “litho” level, new technologies are emerging to tackle challenges in 2.5D and 3D HI, which in turn provide more opportunities for the Hong Kong IC industry.

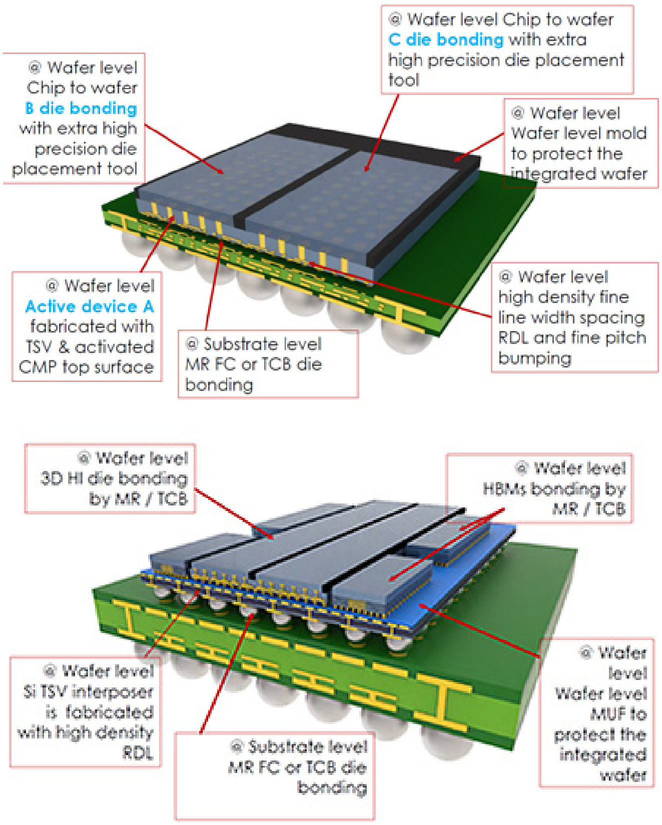

@ Wafer level Chip to wafer B die bonding with extra high precision die placement tool
@ Wafer level Chip to wafer C die bonding with extra high precision die placement tool
@ Wafer level Wafer level mold to protect the integrated wafer
@ Wafer level Active device A fabricated with TSV & activated CMP top surface
@ Substrate level MR FC or TCB die bonding
@ Wafer level high density fine line width spacing RDL and fine pitch bumping
@ Wafer level 3D HI die bonding by MR / TCB
@ Wafer level HBMs bonding by MR / TCB
@ Wafer level Si TSV interposer is fabricated with high density RDL
@ Substrate level MR FC or TCB die bonding
@ Wafer level Wafer level MUF to protect the integrated wafer

Figure K7-4. 3D HI and how it combines with other technologies

Keynote 8:

Discontinuities Drive 3DIC & Chip Integration

Dr. Walden Rhines, CEO Emeritus, Mentor, a Siemens Business

Keynote 8: Discontinuities Drive 3DIC & Chip Integration

Dr. Walden C. Rhines, CEO Emeritus, Mentor, a Siemens Business gave his views on the development of the integrated circuit (IC) industry from a packaging viewpoint captured under the headings below.

MOORE'S LAW COMING TO AN END?

It seems impossible to make transistors any smaller, even Gordon Moore has said "No exponential is forever" in 2003 at ISSCC, meaning it just cannot go on. However, the learning curve inspires us. According to the learning curve plotted on a log-log scale from 1954 to present, the growth of transistor unit volume has sustained around 33% per year cost reduction. Though it is slow, the cost per transistor will always come down. Besides, 3D FLASH memory keeps the overall transistor learning curve on-trend. Therefore, memory is dominating the architecture. Moreover, it can be observed that the cost per function will continue to decline long after Moore's law is obsolete. It is easy to reduce the cost per transistor by shrinking the feature sizes or growing the way for diameters. But someday this will be over and maybe in the near future, the silicon transistors will be replaced by bio switches or spintronics, or something else.

It is remarkable when we look back at the history of the semiconductor industry, the revenue per unit area has been basically constant since germanium. The reason why the revenue per unit area can be constant is when we shrink transistors typically at 70% linear shrink, we double the number of transistors, then we give the customers twice the transistors and charge a little bit more from them. Consequently, we know the cost per unit area has been constant and expect it will continue.

Dr. Walden Rhines, CEO Emeritus, Mentor, a Siemens Business

CONTINUED INCREASE IN ELECTRONIC FUNCTIONS PER UNIT AREA REQUIRES PACKAGING INNOVATION

Since we are not shrinking the transistors anymore, it is a problem with a continued increase in the electronic function per unit area. This is going to require packaging innovation.

There are various integration approaches. For instance, it can be 2.5D (e.g., silicon interposers, thinned silicon "bridge" interposers, and laminate or package interposers), die-to-die (e.g. TSVs like HBM, flip-chip, and direct bonding), or the most promising one: chiplets on silicon with or without substrate. It is also worth noting that the above methods are not mutually exclusive choices and can be used as a combination. From Intel's heterogeneous packaging example, people start putting more functionalities per unit area and the revenue per unit area continues at roughly the same pace it has always been.

Chiplets and heterogeneous multi-chip packaging is an alternative to Moore's law limitations and the integration of disparate technologies. It is an efficient integration of analog, digital, RF, MEMs, etc. for IoT. Additionally, its AI performance requires faster memory access time and a reduced form factor. However, chip integration may be more expensive per square mm than DIS-integration. Therefore, if the processes are much cheaper or if the ecology lends itself to a different type of implementation, the most cost-effective solution can be the multi-chip solution.

FROM MULTI-CHIP LAMINATE TO FAN OUT WAFER LEVEL PACKAGING

Packaging engineers deal with laminate silicon, while silicon engineers deal with silicon. A revolution is happening in the industry, where the assets are now adopting silicon. At present, there is a trade-off that we can let silicon wafer foundries do packaging and make products. Assume we have a typical IoT device, and there are multiple chips on it. It is on a polymer laminate and they have to be put together, designed and verified. However, as we go forward, there will be more and more mixtures of those multiple devices or triplets on a silicon substrate or other implementations.

Different substrate materials dictate different design flows. When we communicate with a packaging engineer, we will find the layout of a package is an output in Gerber or ODB++, no matter if it is ceramic or laminate. On the other hand, engineers in silicon talk about GDSII. The PCB and packaging people use Windows, and the semiconductor people use Linux. The design drives manufacturing in a typical OSAT (outsourced semiconductor assembly and testing). They will set rules that are followed exactly when you show them the multi-chip solution you want and the process flow.

BRIDGING THE PACKAGING AND CHIP DESIGN WORLDS

Bridging the packaging and the chip design world is part of this evolution. We will keep making packaging more efficient and getting more functionalities per unit area. Shapes are domain-specific and minimizing stress in packaging requires curvilinear features. However, forcing curved structures on a gridded system brings errors. For instance, GDSII is all polygons, so it breaks up into straight lines. Each one of those is a design rule checking (DRC) rule violation since there are no curved lines ingenious. Today, there are no curved lines in GDSII, it is all Manhattan rectilinear, so it does not translate very well. If you have a layout that contains curves, then the grid snapping and other things produce artifacts in the silicon layout.

In some real artifacts, there are short or mystery spikes one cannot explain. All sorts of stuff like this occurs regularly. Fixing those problems by editing the GDSII by hand suffers the risk of bringing errors into the design. Additionally, there is no way to go back to the original layout data, once you start editing GDSII by hand. Therefore, hand-edited GDSII is very tough to manage.

Fortunately, the EDA industry has solved this problem by putting the environments together. It bridges between two different operating systems, Windows and Linux. In other words, it allows engineers to go through the silicon world of rectilinear Manhattan labs and the PCB and packaging world, and all these translations are automatic. Therefore, those errors will not occur anymore.

IC STYLE LAYOUT ALSO BRINGS WITH IT ASSUMPTIONS ABOUT LAYER DEPTHS

Unfortunately, that is not the only issue here. When we are going between the packaging world and the silicon world, there is a problem with the layer depths. When we have chips of different layer depths, we have to do some layer matching, or we do not have continuity of the circuit design. What we have to do is to create a PDK that recognizes each process differently, identifies the layers, and then the interposer can connect the right layers, and finally do the verification. You can check and trace the connectivity and compare the source netlist from the package planning environment.

ENABLING A VERIFICATION METHODOLOGY WITH MINIMAL ECOSYSTEM DISRUPTION

Enabling a verification methodology has evolved pretty slowly over time, but there are more problems on the verification side. We need to think about how we could verify the multi-die package was done correctly and the silicon substrate would work. Today, we can get off-the-shelf software that will do all this verification. However, it is in two categories. One is basic physical verification DRC, another is logical verification. Logical verification comes from the silicon world, where logical verification requires an active device. We cannot just run it into an interposer since it should connect to something else. Again, bridging between the packaging world and the silicon world has been done.

BEYOND DRC AND LVS: MODELING – THERMAL, EMI, STRESS TIMING PARASITICS

Once we add all hooks together, we have the problem to check whether all the subtleties will work. Therefore, including the DRC and logic checking, we need to have all things modeled, e.g. thermal analysis, electromagnetic radiation, stress analysis and timing parasitics are required.

The parasitic analysis is to take the parasitics from each of the chips, then to the substrate, and any others you want to consider. It adds them all up and creates dummy devices to collect the parasitics between the devices. Therefore, it can add them up and give you an accurate and reasonable number for the static timing analysis of the pack.

Thermal used to be rule-based. Nowadays, we actually do modeling and simulation. Through thermal analysis, we expect to find how temperature will affect the performance of the part since if we are going to stack memories on top of logic, they are going to be heated in wherever the memory bus goes. The job gets done not as well as it might, and still have a lot of refinement to occur.

As for stress analysis, there are off-the-shelf tools we can buy. However, the stress analysis done by the tools usually ends up in the bumps between dots. Therefore, different thermal coefficients of expansion will appear, which can fracture to or cause other stress problems.

Lastly, we have all of the OSATs of the world, who now see the silicon foundries taking their business. They are moving pretty rapidly to put silicon substrate capabilities in place. The big difference is that a silicon native operation has a process design kit. It is rigid and fixed, and we can get one by asking our vendor. It is not so clear to ask for our assembler for assembly design kit since it varies, different vendors will have different approaches, and sometimes they are negotiating.

COMPONENT TESTING IN 3D

When doing a traditional test, we really do not throw away many components, and the thrown ones only take up a small percentage. Besides, we know they do not cost much. However, when we are going to put all these dies in the same package, yield them, and throw away parts at the end, we are throwing away very expensive parts and the probability that we will have a defect is far greater than we are just processing a single die. Fortunately, IJTAG allows us to connect scan chains between devices, and switch those scan chains to do intermediate testing. The road to chiplets is viable, since the fundamental methodology exists which can be adapted as well.

CONCLUSIONS

Packaging will be a major contributor to future improvements in semiconductor cost, performance, and power. Heterogeneous multi-chip assembly on laminates will progress toward silicon-based approaches including interposers and chiplets. Besides, EDA tools have automated translation between laminate and silicon layouts and have automated rule-deck generation used for physical verification, timing parasitic, EMI, stress and thermal effects. Modeling for multi-dice simulation is improving rapidly but not yet turn-key. Moreover, IJTAG, boundary scan cells, compression, and scan switch network-based testing standards are available. The chiplet libraries and design interface standards are still evolving.

Panel Discussion:

What's kind of microelectronics infrastructure is needed in Hong Kong?

Moderator:

Dr. Mei Kei Ieong, CEO,
United Microelectronics Centre (Hong Kong) Limited

Panellists:

Dr. H.L. Yiu, Head of Advanced Manufacturing,
Hong Kong Science and Technology Parks Corporation

Prof. Tim Cheng, Dean of Engineering,
The Hong Kong University of Science and Technology

Mr. Lincoln Lee, PacRim Technical Director,
Mentor, a Siemens Business

Mr. Wilson Yu, Board Director,
China Resources Microelectronics Ltd.

Mr. Matthew Leung, Director,
Hong Kong Research Centre, Huawei

Mr. Nelson Fan, Vice President,
APT Business Development, ASM Pacific Technology

1. We have been talking about the infrastructure and ecosystem for microelectronics in Hong Kong. What do you and HKSTP have in mind in terms of enabling this infrastructure and ecosystem?

Dr. H.L. Yiu, Head of Advanced Manufacturing of Hong Kong Science and Technology Parks Corporation stated that re-industrialization is one of the Hong Kong Government's initiatives, which is a golden opportunity for us. HKSTP has promoted the development of start-ups and technology companies for over 18 years, where some companies indeed have the potential of commercialization and being strong contenders to existing suppliers of semiconductors. Beyond "soft" infrastructure such as EDA and R&D technologies, "hard" infrastructure is also essential for companies. HKSTP plans to spend two billion Hong Kong dollars to build a microelectronics centre in Yuen Long, covering a space of 30000 square meters with 18000 square meters of cleanroom. With this platform, they are looking forward to collaborating with these technology companies and providing more "content" for the microelectronics industry of Hong Kong.

2. From the perspectives of yourself and ASM Pacific, what do you think of this ecosystem and how will ASM participate?

Mr. Nelson Fan, Vice President, APT Business Development of ASM Pacific Technology thought it is a great opportunity to make things happen and do things right. ASM has developed many technologies and built a lot of tools in Hong Kong. In particular, the high-precision bonding tools like die-bond tools are developed and built in Hong Kong, and over 200 sets of TCB tools have been sold worldwide. ASM has a worldwide interconnect and packaging market. As an international company that has business in Taiwan, Korea and China, they thought the needs and demand are growing. They have set up innovation centres in Hong Kong and mainland China and a R&D centre in Chengdu and Netherland. This is how they work collaboratively with worldwide leaders. In short, ASM is looking for a good platform and source that could enable them to take a step further to grow the ultimate capability.

3. What do you need if we really want to build a platform or infrastructure in Hong Kong so that we can do things together? What do you think you need and how you and your company may participate?

Mr. Matthew Leung, Director of Hong Kong Research Centre, Huawei said their Research Center was established in Hong Kong for 11 years to do chip design and around 180 engineers are now doing solely R&D. He echoed that there are more activities in China going on these days, and Huawei has taken part in some of them, like 2.5D packaging and other advanced stuff. Due to the geopolitical situation, it is even more critical for Chinese companies to have technology ownership. Hong Kong now acts as a bridge between the Western world and mainland China in the field of technology. Huawei built the research centre in Hong Kong because it thinks Hong Kong is an excellent platform, not just for technologies but also talents. Huawei is becoming more and more famous and can attract more talents to come to Hong Kong. However, he thought more companies like Huawei are needed. Once those big companies form a critical mass, and with a stable R&D place, more world-class academics and talents will be willing to come.

4. With the resources in Hong Kong, how to improve the affinity with the inner Hong Kong activities?

Prof. Tim Cheng, Dean of Engineering at The Hong Kong University of Science and Technology thought that the advantage of Hong Kong is that it has multiple excellent universities in the microelectronics area, which can attract talents. He strongly agreed with Dr. Leung that microelectronics and semiconductor are heavily talent-intensive industries. Without talents, no matter how much money and how big a market we have, we are not going to establish an ecosystem. Besides, Hong Kong is a place that protects IP and capital mobility well. Hong Kong can be a perfect research hub for microelectronics. In addition to design, what Hong Kong can do is to develop research prototyping capability, which he called nano system facilities. He mentioned that we should not do silicon foundry which we cannot afford and do not need, what we need is to be able to prototype on a silicon wafer with emerging memory, MEMS, silicon photonics, power electronics combined with III-V compound materials, all these heterogeneous or even monolithic integration. We need this prototyping capability and once we have it, we can let it combine with our world-class IC and EDA centre. This is also where the universities in Hong Kong excel in microelectronics, from materials, process, design, system to application. Furthermore, the Greater Bay Area offers Hong Kong an excellent opportunity to become a part of the ecosystem. Hong Kong is in an up-and-coming region, which can build a complete ecosystem from the user to the manufacturer.

5. Being a veteran in the industry, could you please share with us your opinions in the process and fabrication areas?

Mr. Wilson Yu, Board Director of China Resources Microelectronics Ltd said he joined this industry in the wave of manufacturing since 1983. All the way up here, He has only joined two to three companies, all are process and fabrication related. His first 15 years was in Hong Kong, another 21 years was in Wuxi and Shanghai. From his point of view, even a billion from the Hong Kong Government is not enough. The funding now is just enough to start, we still need to think what else can be done. His feeling is that do not try to repeat those high volume foundries like what is being done in China nowadays. The opportunities in Hong Kong are international talents and great location which is close to the huge market in China. We can solve a lot of problems in Hong Kong and leverage its niche, such as using CMOS wafers from China and talents from Hong Kong and the globe. Mr. Yu used a case in his company about employing Silicon carbide for power electronics, that his company does not have the right tools, so they use the resources and facilities available in Hong Kong to achieve their task.

6. You have been working with a lot of top customers in the world and you also know about Hong Kong. Can you share your opinion about what could be done in Hong Kong, both from EDA and the other perspectives?

Mr. Lincoln Lee, PacRim Technical Director of Mentor, a Siemens Business said he talked with many customers recently when he was attending conference in Taiwan and heard two words a lot, namely, wholesome view and ecosystem. There are two roles that he thought Hong Kong can play. First, Hong Kong can be a place to integrate all pieces to make the whole AI chip industry work, that no single

place can put them together at the moment. The second role Hong Kong can play is a gateway between a lot of IC design companies and the backend manufacturers. Those two parts are now working independently, Hong Kong can then be a gateway to negotiate with both sides and make things work.

7. How to differentiate Hong Kong as a microelectronics infrastructure hub among the places which have strong heterogeneous ecosystem? How to attract more talents, and what approaches are needed to promote Hong Kong in this sense?

Dr. H.L. Yiu expressed affirmation for the opportunities that lie in Hong Kong. Considering demand for technologies, half of the potential customers are from local companies. The reason for choosing Hong Kong instead of other places is that they can control the process, and furthermore they are able to obtain better funding from the investors and get themselves stabilized. The funding can be well utilized to cover the gap of lack of prototyping capability.

Prof. Tim Cheng thought Hong Kong is a part of the Greater Bay Area which possesses a powerful ecosystem. He stressed that the microelectronics industry will attract more investment in Hong Kong than other places due to the matureness of microelectronics in Hong Kong. From the perspective of talents, Hong Kong does have the flexibility and determination to attract talents because there are not only many top universities in Hong Kong, but also companies are offering positions with high salaries.

Mr. Wilson Yu revealed the fact that the projects in China mainly focus on the scale of market rather than the special technologies. However, in Hong Kong, all fundamental elements such as people and the law system are prepared for innovation. If the Hong Kong government can support some centres to have the necessary ecosystem and infrastructures, the small companies will be able to grow and become influential. With such environment and atmosphere, the talents might come from all over the world, such as UK, Europe, Japan and US, etc and feel comfortable in Hong Kong. And the most important thing is that the goal of innovation is not to conquer the whole market, but to be more focused.

Dr. Matthew Leung pointed out the missing part in Hong Kong is the downstream industry. Since there are good researchers in Hong Kong and lots of industries in Shenzhen, the remaining thing is to establish the bridge. Even though they have some collaborations in the past, the level is not deep enough. Based on this, Matthew stated that the government should support more collaboration. In fact, those companies are interested about the talents and technologies that Hong Kong could provide, and the kinds of things they can do in Hong Kong.

LOOKING AHEAD

LOOKING AHEAD

A clear consensus emerged from the panel discussion at the AI Chip Summit 2019 that Hong Kong has the right ingredients – such as strong global talents, the robust rule of law and geographical advantages – to be an ideal base in Asia for AI and microelectronics research and development.

Dr. Walden Rhines, CEO Emeritus of Mentor, a Siemens Business, said he believed that Hong Kong's unique attributes of open communications, well-educated and innovative talents and proximity to China would accelerate next-generation microelectronic development. Mr. Matthew Leung, Director of Hong Kong Research Centre of Huawei, suggested Hong Kong could serve as a bridge to connect technologies and human resources between the West and China. Prof. Tim Cheng, Dean of Engineering at The Hong Kong University of Science and Technology, stated that the excellent universities in Hong Kong provided top-notch talents for the industry, and that strong IP protection and capital mobility made the city a perfect research hub for microelectronics. Mr. Wilson Yu, Board Director of China Resources Microelectronics Ltd, agreed that Hong Kong has huge opportunities in AI chip development, supported by its international talent pool and enviable location at the doorstep of the huge market in mainland China.

Inherent advantages aside, Hong Kong still faces keen competition regionally in the AI R&D race. Japan is putting AI, Big Data and IoT at the centre of the country's Revitalisation Strategy, supported by a budget of US$723 million set aside in fiscal 2018. Singapore has committed US$371 million for AI development in its National Artificial Intelligence Strategy 2019 and will invest more going forward. Taiwan rolled out the AI Taiwan Action Plan in 2018 to accelerate innovation and development with a total budget of over US$1 billion.

To stay competitive in the region, Hong Kong is making a concerted effort to empower AI development while building the city as an international innovation and technology hub. Hong Kong Science and Technology Parks Corporation (HKSTP) is actively facilitating its AI and Robotics technology platform by providing funding, space and facilities, incubation programmes and commercialisation opportunities to support the R&D activities of its extensive community of start-ups and partner companies. UMEC(HK) acts as a key player in the global semiconductor industry supply chain, is determined to drive innovation and advancement of the AI chips and systems. HKSTP and UMEC (HK) will work together to support Hong Kong in the global AI race.

The rapid development of AI has brought new opportunities to enterprises as well as local start-ups. Some emerging technologies outlined in this whitepaper, such as applying machine learning to EDA design tools and 3DIC heterogeneous integration, are creating new opportunities for the industry. Hong Kong can be positioned in two ways to capture these opportunities — one is being a place to integrate every piece of the AI chip industry, and the other is to act as a gateway for all players in the supply chain, including IC design and system companies, EDA vendors, wafer and back-end manufacturers. In addition, the Hong Kong Government could act as a facilitator to bring together industry stakeholders and form a consortium or alliance to tackle various challenges and help achieve its identified value proposition in the global semiconductor industry.

In conclusion, rapid growth is certain for the AI industry, and there is no doubt that the demand for AI chips and systems will increase dramatically in the coming years. This is a golden era for Hong Kong to develop its capabilities in microelectronics and advanced AI technologies and become a vital part of the Greater Bay Area ecosystem. Hong Kong's competitive edge, together with the government support, will engender a promising future for AI technology advancement and industry development that will generate significant benefits for Hong Kong's overall economy.

About Hong Kong Science and Technology Parks Corporation

Comprising Science Park, InnoCentre and Industrial Estates, Hong Kong Science & Technology Parks Corporation (HKSTP) is a statutory body dedicated to building a vibrant innovation and technology ecosystem to connect stakeholders, nurture technology talents, facilitate collaboration, and catalyse innovations to deliver social and economic benefits to Hong Kong and the region.

Established in May 2001, HKSTP has been driving the development of Hong Kong into a regional hub for innovation and growth in several focused clusters including Electronics, Information & Communications Technology, Green Technology, Biomedical Technology, Materials and Precision Engineering. We enable science and technology companies to nurture ideas, innovate and grow, supported by our R&D facilities, infrastructure, and market-led laboratories and technical centres with professional support services. We also offer value added services and comprehensive incubation programmes for technology start-ups to accelerate their growth. Technology businesses benefit from our specialised services and infrastructure at Science Park for applied research and product development; enterprises can find creative design support at InnoCentre; while skill-intensive businesses are served by our three industrial estates at Tai Po, Tseung Kwan O and Yuen Long. More information about HKSTP is available at https://www.hkstp.org

About United Microelectronics Centre (Hong Kong) Limited

United Microelectronics Centre (Hong Kong) Limited is a research and development company in Hong Kong, focusing on AI chips & systems and 5G applications development. We respond to the dynamic market environment by adopting an innovative research centre operation model, highlighted by talent-centric culture and tailored research projects. Through continuous cooperation with global top echelon universities and research institutions, we bring the AI and 5G technologies to the next level and drive commercialisation.

Contact Us:

HKSTP:

☎ +852 2629 1818

✉ **International:**
international@hkstp.org

✉ **Mainland:**
prc@hkstp.org

UMEC(HK):

☎ +852 3643 1606

✉ info@umechk.com

Stay tuned for our upcoming events!