Who is leading the AI chips IP race? Best practice in IP management

By Bruce Rubinger and Jason Hannon

February 12, 2020

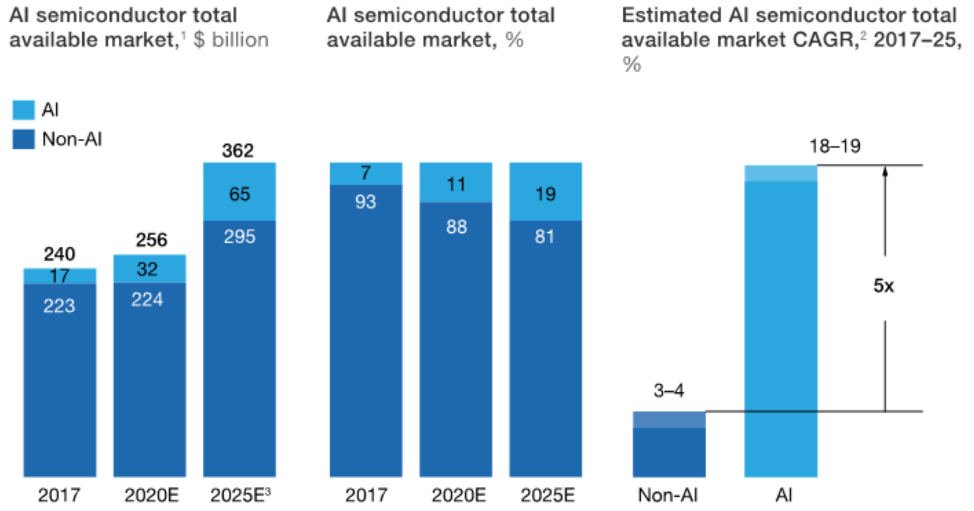
AI and its diverse applications (eg, machine learning and deep learning) have seen significant market growth in the past decade and are on the cusp of transforming many industries. As one of the hottest technologies, AI underpins the performance of:

- data centres;
- voice assistants;
- targeted ads;
- medical diagnosis;
- product development;
- oil prospecting;
- insurance;
- security;
- driverless cars; and
- other needs.

Empowering this transformation are specialised AI chips designed to optimise specific applications. This article highlights the deep insights that can be gleaned in a crowded market from an accurate IP landscape analysis of the opportunities for creating strong IP portfolios and the IP positioning of players in this critical AI-enabling sector.

Areas that use AI for the Cloud range from cloud computing applications to digital assistants, self-driving and autonomous vehicles, and medical diagnosis. AI has also been adopted for advanced product development, as illustrated by the semiconductor industry, where it is employed to address the growing complexity of chip design. It is reported that AI processors implementing neural networks improve performance beyond that of other analytical techniques for diverse applications. These benefits have spurred the development of AI chips that can process data faster and more efficiently.

The semiconductor market for AI-related chips is projected to grow substantially over the next few years. Mckinsey & Company predicts a tripling in revenue for AI chips from $17 billion to $65 billion (see Figure 1). By 2025 such revenue is expected to account for nearly 20% of semiconductor sales. Other studies (eg, by Research and Markets) predict that the global AI chip market will reach $90 billion by 2025, growing at a compound annual growth rate of 45.2% from 2019.



[1] Total available market includes processors, memory, and storage; excludes discretes, optical, and micro-electrical-mechanical systems.

[2] Compound annual growth rate.

[3] E = estimated.

Source: Bernstein; Cisco Systems; Gartner; IC Insights; IHS Markit; Machina Research; McKinsey analysis

Figure 1: Growth for semiconductors related to AI is expected to be five times greater than growth in the remainder of the market

AI usage and processing types

There are essentially two main areas of AI at present: cloud AI and edge AI (see Figure 2). Cloud applications are those where all data is fed to a remote data centre and the processing is done on the Cloud. Working on the Cloud enables businesses to move faster, more efficiently and at a lower cost. Edge computing is important where a lack of speed or connectivity with the Cloud requires processing applications closer to the data source (eg, through a machine on a factory floor, an MRI scanner at a hospital or advanced phones).

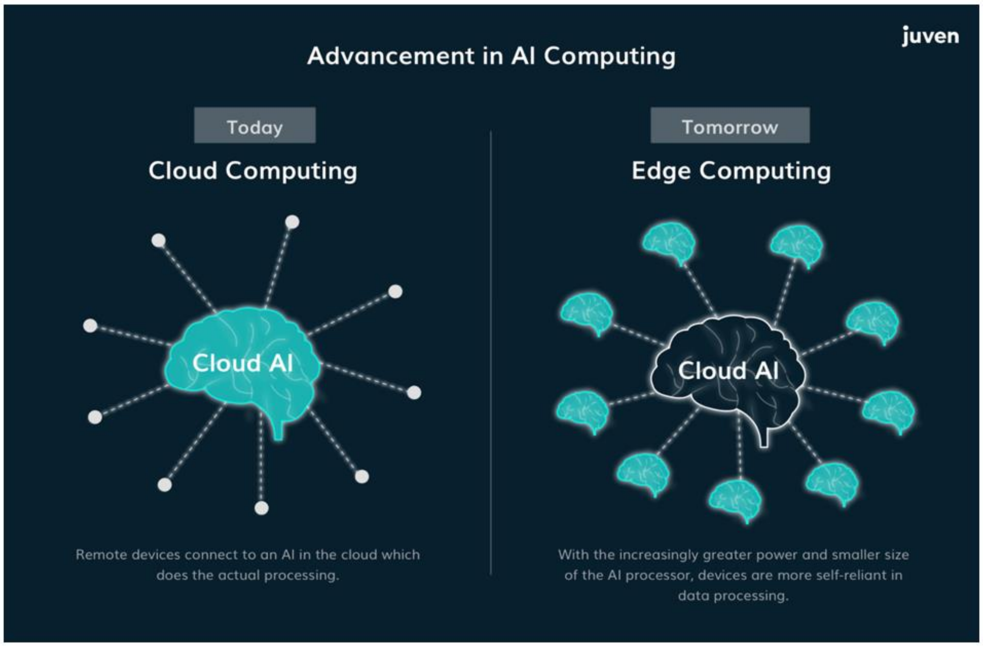
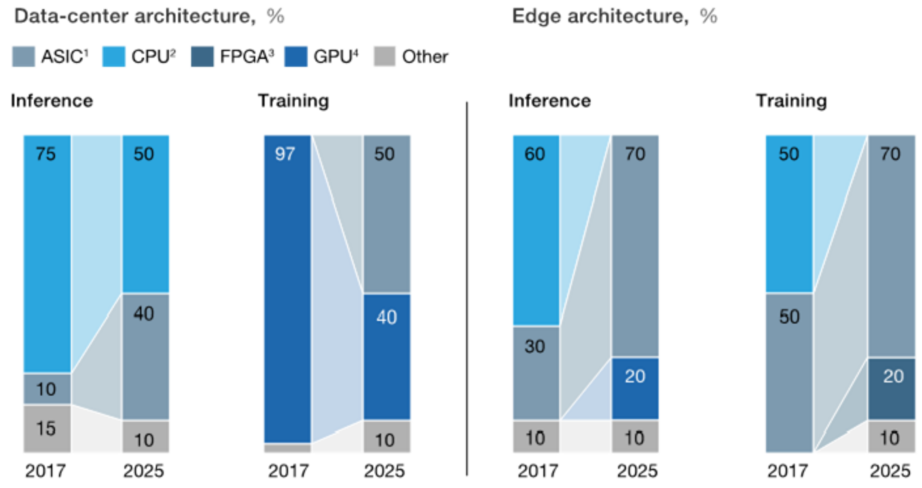

juven
Advancement in AI Computing
Today
Cloud Computing
Cloud AI
Remote devices connect to an AI in the cloud which does the actual processing.
Tomorrow
Edge Computing
Cloud AI
With the increasingly greater power and smaller size of the AI processor, devices are more self-reliant in data processing.

Figure 2: Cloud and edge AI

Cloud AI

Currently, for cloud-based AI (data centres), most computing is provided by central processing units (CPUs) and graphics processing units (GPUs). However, a major shift in preferred chip architecture is underway to meet the performance needs of AI computing (see Figure 3). GPUs have traditionally been attractive for implementing neural networks since image processing requires parallel tasks involving matrices, which is efficiently addressed by neural networks. In contrast, traditional CPUs can be programmed to conduct AI tasks but take longer and use more power for the same process. Several studies (eg, McKinsey) predict a significant growth in application-specific integrated circuits (ASICs) (see Figure 3). The patent data confirms that semiconductor companies such as IBM, Intel, and Qualcomm are designing ASICs to improve power efficiency and increase throughput and the technical path that they are pursuing. These firms are also focused on developing AI chips that can be trained efficiently – the step that entails preparing a machine-learning model by feeding it data from which it can learn. Inference is the process of taking a model that has already been trained and using it to make useful predictions.



[1] Application-specific integrated circuit.
[2] Central processing unit.
[3] Field programmable gate array.
[4] Graphics-processing unit.

Figure 3: The preferred architectures for compute are shifting in data centres and the edge
Expert interviews; McKinsey&Company analysis

End of traditional processor company dominance?

AI represents a game-changing technology that could end the dominance of traditional processor design companies. Initially, AI was performed by CPUs, before then moving to GPUs, as these were better suited for parallel processing. While GPUs still excel at dense floating-point computation, researchers have reported higher throughput and energy efficiency with custom hardware. The patent analysis found that a significant number of IT firms have selected custom hardware over CPUs for implementation of their neural network architecture. Customisation of integrated circuit logic and memory hierarchy can yield custom hardware neural networks that are faster and significantly more energy efficient than the previous generation of GPUs.

International competition has transformed this IP space. In 2017 China unveiled its Next Generation Artificial Intelligence Development Plan, a document that outlined the country’s strategy to become the global leader in AI by 2030. Leading Chinese tech companies such as Huawei have started to design and file intellectual property in AI. Indeed, Huawei has announced an AI core for a system on a chip used in its phones. Search giant Alibaba is another new entrant in AI chip design. Horizon Robotics is focused on the design of AI chips for surveillance cameras as well as for autonomous vehicles.

In the United States, traditional semiconductor firms such as Intel, IBM, Qualcomm, AMD and NVIDIA have either announced or already shipped cloud AI chips. In addition, non-traditional semiconductor companies seeking to enhance their position in cloud computing (eg, Google, Microsoft and Amazon) have invested significant money in developing AI chips for the cloud.

Announced chips

Figure 4 presents a list of companies that have announced AI chips for cloud computing. The data shows that the United States and China are leading the race, with 16 active companies. In contrast, only four players from the rest of the world (Europe, Israel, Japan and South Korea) have announced AI chips. Also significant is the number of players that lack traditional chip design expertise, ranging from start-ups (Cerebras, Graphcore, Canaan, Cambricon) to data companies (Google, Baidu, Alibaba).

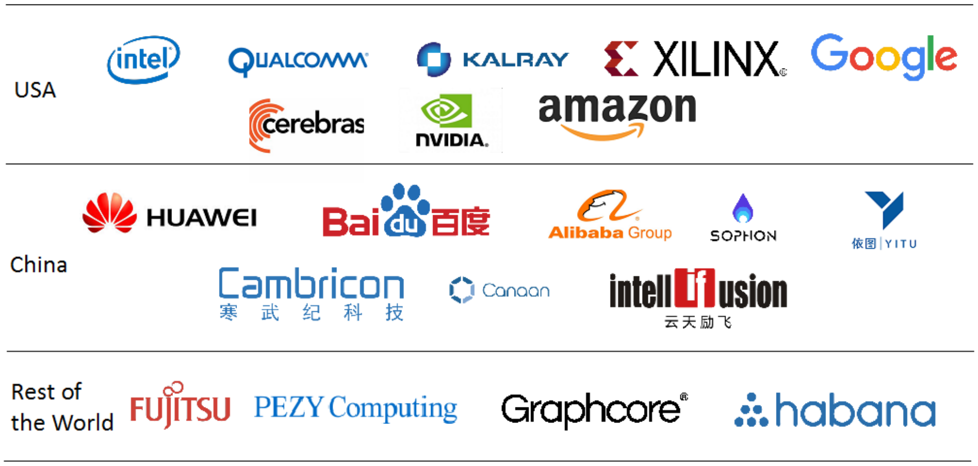


Figure 4: Companies that have announced AI cloud chips

Figure 5 shows the significant increase in AI chip announcements each year. For example, 2019 had double the number of chip announcements of 2018. This trend will likely continue, as companies that have not yet announced chips (eg, IBM, AMD, Microsoft and Facebook) will announce their products in the coming months, while those that have already announced their chips will continue to bring out new designs.

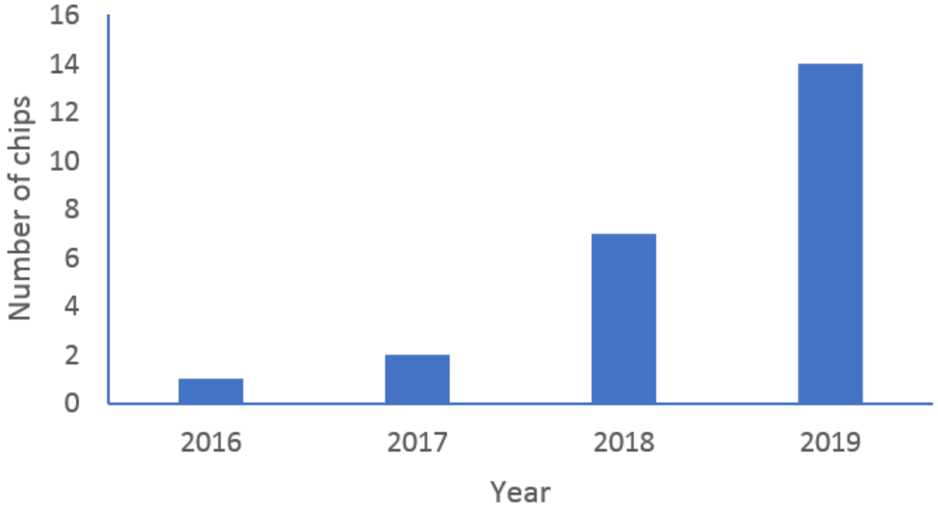


Figure 5: AI chips announced by year.
GPA Research

AI chip patent holders and methodology

The search process was carried out manually for accuracy, in order to avoid flagging patents that were not relevant but which used terminology that might overlap that of AI chips (noise), while identifying important patents where the design and intended application were apparent from the exhibits. Given the complex technology, this process was conducted by Global Prior Art Inc (GPA) technical experts in AI chip technology, who examined the chip implementation and associated teachings. The research identified patent portfolios for multiple companies, including established players such as Intel, IBM, Qualcomm and Nvidia, as well as start-ups such as Graphcore, Cerebras, Habana Labs and YITU. In addition, portfolios filed by new entrants that lacked a traditional semiconductor design background were identified; this group includes Amazon, Google, Baidu and Alibaba. The search yielded more than 2,000 distinct patents families with most patents held by traditional semiconductor firms.

For Chinese companies, GPA's semiconductor specialist searched native-language Chinese patents, uncovering highly relevant Chinese patents that lacked a US counterpart. This is illustrated by Intellifusion, which filed 246 distinct Chinese patent documents, of which there are only two US counterparts. The body of 2,000 distinct patent families relating to AI chips was further analysed with regard to technology, intended application and innovative focus for the case studies.

A case study comparison of the AI chips patent landscape

Three case studies have been carried out to highlight the IP landscape for AI chips designed for cloud computing. The first study compares the two incumbent processor companies, Intel and IBM – traditional semiconductor companies with extensive technical expertise and large IP portfolios. The second compares the patent portfolios of AI chip newcomers Qualcomm and Cambricon, both of which have produced an AI processor for mobile uses and have announced plans to pursue AI chips intended for cloud AI. The final case study compares the IP portfolios of Google and Baidu – two cloud companies that are actively developing AI application-specific integrated circuits (ASICs).

Figure 6 shows the number of distinct patent families that selected companies have filed over the past five years. What is evident is that traditional semiconductor companies have been patenting at a high rate. Indeed, Intel has 160 patents, followed by IBM with 82 distinct patent families in this period. Cambricon, which was founded in only 2016 has more than 60 patent documents while Google's portfolio of 34 patents demonstrates how serious it is about entering this space. The most active filer was Intellifusion, a recent entrant in the AI Cloud Chips field based on China that filed more than 250 patents, the majority of which were filed only in China (246), while the remaining few are split between Patent Cooperation Treaty (PCT) patents and US filings.

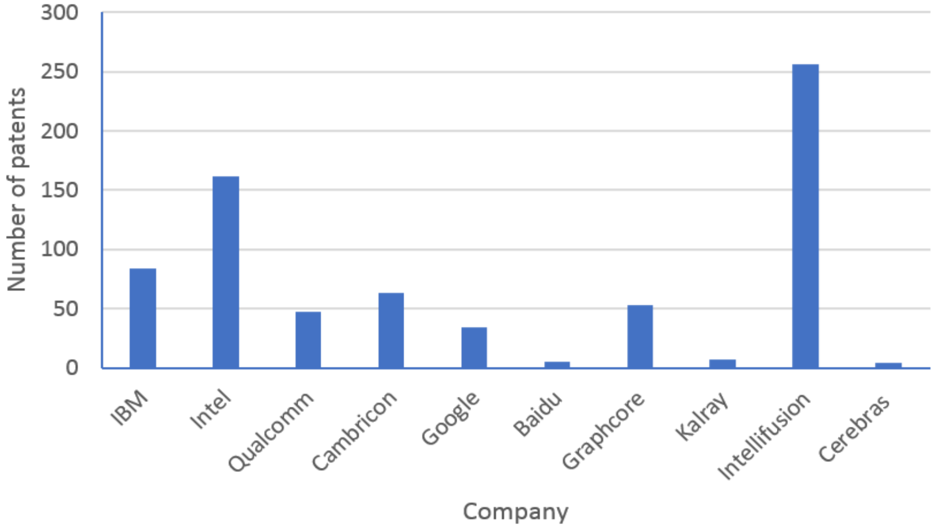


Figure 6: Number of distinct patent families over the past five years

Figure 7 provides a breakdown of filing trends for these companies by year. Intellifusion and Graphcore have both filed the majority of their AI chip patents in the past three years. Meanwhile, for Intel, 2017 was a big year and IBM shows a sustained effort over the entire five-year period.

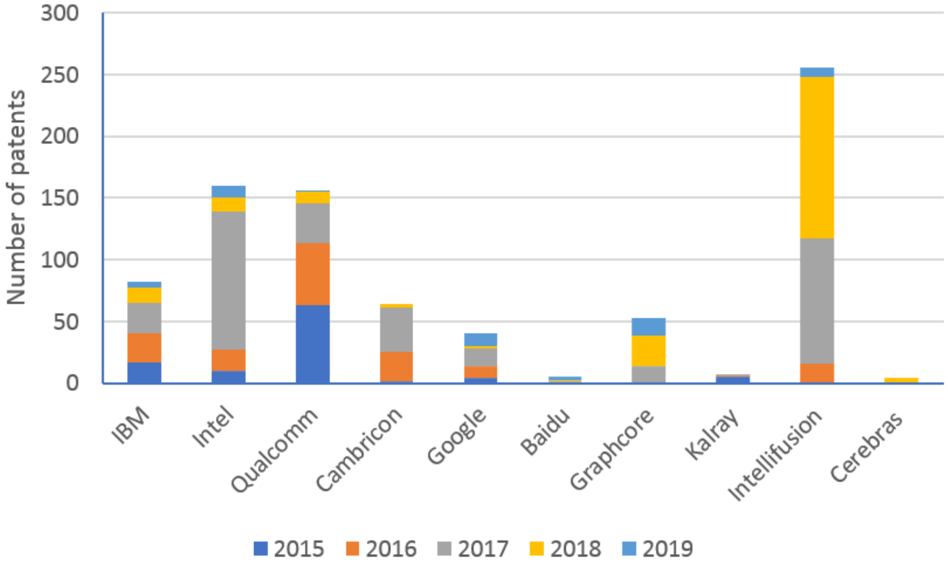


Figure 7: Number of patents filed by year

Case study one: Intel and IBM

The two semiconductor behemoths traditionally supplying server chips are Intel and IBM. With demand for server chips specifically designed for AI growing rapidly, it is strategically important for these companies to develop AI intellectual property and chips designed for the cloud, without ceding market share to newcomers.

IBM has actively researched AI for decades and, more recently, made a multibillion-dollar investment in growing its Watson business, pursuing diverse applications of AI technologies that span medical diagnostics, manufacturing and insurance, among other things.

Intel has also conducted AI-related research for decades, while simultaneously acquiring companies that develop AI chips in order to capture the market. In 2016 it acquired Nervana systems – developing the Nervana engine optimised for deep learning, which reportedly performs 10 times better than graphics processing units (GPUs) – and Movidius, , which designs specialised low-power processor chips for computer vision. In 2017 it then acquired Mobileye, which had been developing advanced driver-assistance systems for collision prevention. As a result of these takeovers, Intel recently introduced two new Nervana neural network processors – its first ASICs designed explicitly for AI in the cloud. The NNP-T is designed for training, while the NNP-I is designed for inference.

IBM meanwhile has announced plans to introduce AI processors that will include both digital AI and analogue AI cores. The company has claimed that the new analogue cores double the precision of previous analogue cores while consuming 33 times less energy than a digital architecture with the equivalent precision. A recent report predicted that analogue techniques may yield up to four times the performance while using three to four times less power than digital techniques (Dahad, 2019).

In total, we identified nearly 1,000 IBM patents relating to AI chips over the past 20 years. For Intel, we identified 558.

Filing trends from the past five years

Taking a closer look at these figures, over the past five years, Intel has filed approximately 180 US patents and IBM has filed 75. A snapshot of the period will help to identify what each company views as important aspects of AI chips moving forward and the associated intellectual property.

Intel snapshot

Intel's filings appear to be strongly driven by the GPU chip Xe, which is due for release in 2021. The chip is being marketed as designed for data centre AI and high-performance computing workloads. It shows Intel's focus on digital implementation techniques, as well as analogue techniques.

An example of this is US Patent 10,102,609, describing a parallel processor coupled to a memory hub through a communications link. The processors form a vector processing system that can include a large number of processing cores. This system is described as being able to accelerate machine-learning operations, pattern analysis operations and various general-purpose functions.

Another example is US Patent 10,261,903, which describes a method to increase processing efficiency for parallel processing systems. It utilises a single instruction, multiple thread architecture designed to maximise the amount of parallel processing in the processing pipeline. The computations required for training and using machine-learning algorithms lend themselves naturally to efficient parallel implementations.

US Patent 10,360,496 describes a digital neuromorphic processor – an AI chip designed to implement models of neural systems for perception, motor control or multisensory integration. In this design, each neurosynaptic core comprises a plurality of neurons with a synapse array comprising a plurality of synapses to communicate with the neurons.

Intel's recent patents highlight the company's focus on a digital approach covering diverse parallel processing and vector processing techniques. Other patents in its portfolio cover memory and communications means for parallel processing. In contrast, before 2015 Intel prioritised analogue techniques.

IBM snapshot

IBM's IP filings over the past five years have been extremely diverse, with 183 patents representing 82 unique patent families. Of these filings, approximately 20 patents focused on analogue AI cores. Other areas of interest included digital implementations, connections and memory. Some of the most noteworthy patents prioritised analogue implementations for AI – an approach that promises faster throughput at lower power usage and may emerge as the long-term winner for AI chips.

One example of an analogue AI circuit is illustrated in US Patent 10,134,472, which describes a resistive processing unit that can be used in neural networks. Other recent examples include US Patents 10,192,161 and 10,381,074.

Case study two: Qualcomm and Cambricon

Qualcomm is well known as a leading designer of processors that are found in most smartphones. The firm now provides AI solutions for smartphones, including object classification for smart cameras. Over recent years, Qualcomm has pushed towards developing AI chips for the data centre and has been actively patenting to protect this effort. This culminated in its announcement of the Cloud A1 100 server chip in April 2019. The main application for this chip is to facilitate decisions based on digital voice or picture data stream analysis. Qualcomm claimed that the chip's AI processing capacity is 50 times that of its mobile phone chip. Production is due to start in 2020.

Cambricon's first AI designs were for smartphone usage and its IP rights for the Kirin 970 chip were licensed by Huawei. Similar to Qualcomm, Cambricon has since focused on the data centre, for which it provides AI acceleration chip solutions. The AI acceleration chip is designed as a hardware acceleration for AI applications – in particular, for artificial neural networks, machine vision and machine-learning applications. In 2018 Cambricon launched its first AI server chip, the MLU100. In 2019 it announced its second-generation AI server chip, the MLU270. Since then, the company has added a video decoding unit to the MLU270, which is intended for the video processing market.

Trends from the past five years

Referring back to Figures 1 and 2, Cambricon has filed 63 unique patent families over the past five years, including filings in China, the United States, the PCT and Europe. During the same period, Qualcomm filed 34 distinct patent families related to AI chips in the United States. Several representative patent documents are highlighted below to illustrate the focus of these firms.

Qualcomm snapshot

Qualcomm's focus is illustrated by US Patent 10,083,378, which teaches the application of deep-learning techniques to video analysis. The implementation describes a processor configured to learn a baseline pattern using deep-learning processing. The processor is also configured to autonomously detect a change due to movement within a field of vision using deep-learning processing.

In US Patent 10,061,909, Qualcomm's focus is on a convolutional neural network (CNN) that learns connections between consecutive samples to predict outcomes based on the extracted features. The CNN's implementations achieve better generalisation on vision problems. In general, Qualcomm's patent filings focus on high-level representations (eg, the '909 patent, where a neural network is described).

Cambricon snapshot

Cambricon's filings tend to focus on high-level patents. WO2019165946 describes an integrated circuit with a main processor and multiple basic processing units where the basic processors perform a parallel data operation before returning the data to the main processor. Meanwhile, Patent Application WO2018112699 describes an artificial neural network with reverse training. Each generation of training adjusts the data according to the learning rate of the previous generation.

Case study three: Google and Baidu

Google is widely known for its software prowess, but in the past few years, it has entered the chip design space. In 2016 the company announced the tensor processing unit (TPU) and in 2017 a second-generation TPU chip with increased memory bandwidth was revealed. The third generation of TPU – which claims to be twice as powerful as the second – was then announced in 2018. TPUs are being used for machine-learning applications such as neural networks. These chips are protected by more than 34 patent families and have been deployed in Google data centres.

Like Google, Baidu is better known for its software than its hardware. However, it has also branched into developing its own chips for its server farms. In 2019 Baidu announced the Kunlun chip, which focuses on the optimisation of visual, speech and natural language processing. Media reports claim that the chip is 30 times faster than field-programmable gate array-based AI accelerators.

Trends from the past five years

Over the past five years, Baidu has filed the lowest number of patents among the firms highlighted, with five patents filed in China, the United States, the PCT and Europe. In contrast, Google has filed 34 patents related to AI chips, in the areas of neural networks, memory and vectors. These patents focus on connections between neural networks, parallel processing and neural network techniques. Representative Google patents include US Patent 9,928,460, which describes a 3D neural network accelerator, and US Patent 10,037,490.

Baidu

Baidu's patents focus on various types of artificial network and how data is handled in a matrix format. Representative patents include US Patent 2017/0169326, which describes a recurrent neural network – a type of artificial neural network where the outputs from units in a given timestep are fed into the inputs of the same units in the next timestep, giving it an element of memory – and US Patent document 2018/0032336 – an implementation which improves the execution efficiency of deep learning.

Overall analysis

Analysis of the IP portfolios and their positioning in the AI chips space shows that innovation is being driven by new technologies, the entry of new competitors into the market, the pursuit of diverse applications and various approaches to attain higher processor speeds and lower power. Traditional semiconductor processor design companies such as Intel and IBM have created large patent portfolios that highlight contrasting priorities and innovation focus. They compete against other semiconductor firms such as AMD, NVIDIA, and Qualcomm while facing challenges from new competitors in the space, including non-traditional semiconductor companies seeking to enhance their position in cloud computing (eg, Google, Microsoft, Graphcore, Cerebras, Cambricon, and Amazon).

China’s strategy to become a leading player in AI chip design and the path to achieving this are evident from the filings by Huawei, Intellifusion, Baidu and Alibaba.

Lastly, the patent data highlights a shift from CPUs to custom hardware, the distinctive innovation focus of each player and the opportunities for corporate IP executives to capture these strategic technologies with strong patents while guiding internal product development efforts through the patent thicket.

Best practice in IP management

Every firm operates in a crowded space, making it critical to have an in-depth understanding of competitors, underlying technologies, the innovation focus and positioning of all players, and the subsequent space for opportunity. The challenge in achieving this is illustrated by the AI chips landscape, which comprises more than 2,000 unique patent families.

Effective IP management strategies require strong claims that capture key technologies and product features. The best IP managers will realise that substantial efforts are necessary to analyse 2,000 patents at the level required for an effective IP strategy, but the benefits of a strong portfolio are significant. Operating blindly yields poor-quality patents that fail to protect a firm’s investment in innovation.

Cursory knowledge of an IP space gleaned from a high-level assessment of the number of patents filed by various players or the number of times that a patent is cited is easy to obtain but masks the critical features that drive an effective IP patent strategy – namely, knowledge of the underlying technologies and implementation, the product features addressed, the intended benefits, claim coverage and how a patent is distinguished over the art. While such analysis is hard work, it provides the necessary insight into which opportunities a firm should prioritise and the detailed gap analysis to support the creation of strong claims. The benefits include critical intelligence on new players and their focus, the foresight to avoid being blindsided, a roadmap for systematically developing innovative products and claims that capture the benefits of innovation.

This knowledge-based approach requires that any IP landscape analysis be accurate and actionable. The following checklist will allow executives managing corporate IP portfolios to confirm that a body of IP landscape research is actually comprehensive and therefore can be relied on for decision making:

- Precision – 100%. Every flagged patent is relevant.
- Recall – 100%. Every relevant patent was picked up.
- Actionable – within 45 minutes an IP manager or board member will know:
    - what sub-set of patents is relevant to their efforts and requires attention;
    - which companies are active in that space;
    - how these competitors are positioned by technology, product features and innovative focus; and
    - the opportunity space on which to focus.

Problems with low precision are common in IP landscapes that are generated by relying on keywords or semantic searching. The result is noise or large numbers of irrelevant patents using similar terminology. Low recall is harder to spot without checking for patents from known competitors. Attaining high accuracy requires patents to be vetted by a technical expert. The benefits from an accurate IP landscape are enormous and enable a firm to take its IP strategy and patents creation process to a much higher level.