## Intel® Nervana™ Neural Network Processors (NNP) Redefine AI Silicon 英特尔神经网络处理器（NNP）重新定义人工智能硅

原文链接：  
原文链接：

As our Intel CEO Brian Krzanich discussed earlier today at Wall Street Journal’s , Intel will soon be shipping the world’s first family of processors designed from the ground up for artificial intelligence (AI): the family (formerly known as “Lake Crest”). This family of processors is over 3 years in the making, and on behalf of the team building it, I’d like to share a bit more insight on the motivation and design behind the world’s first neural network processor.  
正如英特尔首席执行官布赖恩·克扎尼奇（Brian Krzanich）今天早些时候在《华尔街日报》（Wall Street Journal）上所讨论的，英特尔将很快推出世界上第一个为人工智能（AI）而从头设计的处理器家族：这个家族（以前被称为“Lake Crest”）。这一系列的处理器已经有3年多的历史了，我代表构建它的团队，就世界上第一个神经网络处理器背后的动机和设计，与大家分享更多的见解。

Machine Learning and Deep Learning are quickly emerging as the most important computational workloads of our time. These methods allow us extract meaningful insights from data. We’ve been listening to our customers and applying changes to Intel’s silicon portfolio to deliver superior Machine Learning performance. and are powering the vast majority of general purpose Machine Learning and inference workloads for businesses today. We continue to optimize these product lines to support our customers’ evolving data processing needs. The computational needs of Deep Learning have uncovered the need for new thinking around the hardware required to support AI computations. We have responded to this by listening to the silicon and designing a new chip for Deep Learning called the Intel® Nervana™ Neural Network Processor (Intel® Nervana™ NNP).  
机器学习和深度学习正迅速成为我们这个时代最重要的计算工作。这些方法允许我们从数据中提取有意义的见解。我们一直在倾听客户的意见，并对英特尔的硅产品组合进行调整，以提供卓越的机器学习性能。并且为当今企业的绝大多数通用机器学习和推理工作提供动力。我们继续优化这些产品线，以支持客户不断变化的数据处理需求。深度学习的计算需求揭示了围绕支持人工智能计算所需硬件进行新思考的必要性。对此，我们听取了硅的意见，并设计了一种新的用于深度学习的芯片，称为英特尔神经网络处理器（英特尔神经网络处理器）。

The Intel Nervana NNP is a purpose built architecture for deep learning. The goal of this new architecture is to provide the needed flexibility to support all deep learning primitives while making core hardware components as efficient as possible.  
英特尔神经网络NNP是一个专门为深度学习而构建的架构。这种新架构的目标是提供所需的灵活性，以支持所有深度学习原语，同时使核心硬件组件尽可能高效。

We designed the Intel Nervana NNP to free us from the limitations imposed by existing hardware, which wasn’t explicitly designed for AI.  
我们设计了英特尔NelvaNNP来摆脱现有硬件所施加的限制，而这些硬件并不是为人工智能设计的。

### New memory architecture designed for maximizing utilization of silicon computation 为最大化硅计算利用而设计的新存储器体系结构

Matrix multiplication and convolutions are a couple of the important primitives at the heart of Deep Learning. These computations are different from general purpose workloads since the operations and data movements are largely known a priori. For this reason, the Intel Nervana NNP does not have a standard cache hierarchy and on-chip memory is managed by software directly. Better memory management enables the chip to achieve high levels of utilization of the massive amount of compute on each die. This translates to achieving faster training time for Deep Learning models.  
矩阵乘法和卷积是深度学习的核心部分。这些计算不同于一般用途的工作负载，因为操作和数据移动在很大程度上是先验的。因此，英特尔Nervana NNP没有标准的缓存层次结构，片上内存由软件直接管理。更好的内存管理使芯片能够在每个芯片上实现大量计算的高利用率。这意味着为深度学习模型实现更快的培训时间。

### Achieve new level of scalability AI models 实现新级别的可扩展性人工智能模型

Designed with high speed on- and off-chip interconnects, the Intel Nervana NNP enables massive bi-directional data transfer. A stated design goal was to achieve true model parallelism where neural network parameters are distributed across multiple chips. This makes multiple chips act as one large virtual chip that can accommodate larger models, allowing customers to capture more insight from their data.  
Intel Nervana NNP采用高速片内和片外互连设计，实现大规模双向数据传输。一个既定的设计目标是实现真正的模型并行，其中神经网络参数分布在多个芯片上。这使得多个芯片充当一个大型虚拟芯片，可以容纳更大的模型，允许客户从他们的数据中获取更多的洞察力。

### High degree of numerical parallelism: Flexpoint 高度的数值并行度：Flexpoint

Neural network computations on a single chip are largely constrained by power and memory bandwidth. To achieve higher degrees of throughput for neural network workloads, in addition to the above memory innovations, we have invented a new numeric format called Flexpoint. Flexpoint allows scalar computations to be implemented as fixed-point multiplications and additions while allowing for large dynamic range using a shared exponent. Since each circuit is smaller, this results in a vast increase in parallelism on a die while simultaneously decreasing power per computation.  
单芯片上的神经网络计算很大程度上受到功耗和存储带宽的限制。为了提高神经网络工作负载的吞吐量，除了上述内存创新之外，我们还发明了一种新的数字格式Flexpoint。Flexpoint允许标量计算作为定点乘法和加法实现，同时允许使用共享指数的大动态范围。由于每个电路都较小，这会导致模具上的并行度大幅增加，同时降低每次计算的功耗。

### Meaningful performance 有意义的表现

The current AI revolution is actually a computational evolution. Intel has been at the heart of advancing the limits of computation since the invention of the integrated circuit. We have early partners in industry and research who are walking with us on this journey to make the first commercially neural network processor impactful for every industry. We have a product roadmap that puts us on track to exceed the goal we set last year to achieve by 2020.  
当前的人工智能革命实际上是一种计算进化。自从集成电路发明以来，英特尔一直是提高计算极限的核心。我们在工业和研究领域有早期的合作伙伴，他们正与我们一起踏上这一征程，使第一款商用神经网络处理器对每个行业都具有影响力。我们有一个产品路线图，使我们有望超过我们去年设定的到2020年实现的目标。

In designing the Intel Nervana NNP family, Intel is once again listening to the silicon for cues on how to make it best for our customers’ newest challenges. Additionally, we are thrilled to have Facebook\* in close collaboration sharing their technical insights as we bring this new generation of AI hardware to market. Our hope is to open up the possibilities for a new class of AI applications that are limited only by our imaginations.  
在设计英特尔Nervana NNP系列产品时，英特尔再次聆听矽的声音，以了解如何使其最适合客户的最新挑战。此外，我们很高兴Facebook\*与我们密切合作，分享他们的技术见解，因为我们将这一代人工智能硬件推向市场。我们的希望是为一类新的人工智能应用开辟可能性，而这类应用只受到我们想象力的限制。

We hope you’ll join us on this exciting journey to build the future of Artificial Intelligence.