|  |  |
| --- | --- |
| Слайд 1  My name is Ivan Ivanov, I study at the InfoSec department of MIET and present a paper on AI Enhanced Processors.  Слайд 2  This work purpose is to cover the essence of the dedicated Artificial Intelligence engine in modern computer processors.  Throughout the presentation we will discuss such problems as the Moore’s law, processing limits, what are the AI enhanced processors and what are the benefits of having a purpose-built AI accelerator.  Слайд 3  The Moore’s law claims that the number of transistors on an integrated circuit will double every two years.  According to Intel data, gathered over the last 50 years, they forecast that the rate of transistors count growth will keep up just as the Moore’s law predicts.  But physically we are getting so close to the dead-end size of the single transistor, so that it could not be possible to push forward the development of much denser arrangement. This leads to the fact, that we might use some other techniques and technologies to increase the processing limits.  Слайд 4  This leads us to the obvious question: what are the other possible ways for us or manufacturers to push processing limits further?  There were several ways how manufacturers tried and actually increased the processing speeds.  The simplest of them – making a superscalar processors (bigger than the ordinary once).  The Reduced Instruction Set Computing (a.k.a. RISC) was also implemented (worth mentioning ARM processors).  Multi-core and multi-threading were designed to increase overall processing capacity and efficiency.  Resolving a power consumption and cooling problems can lead to frequency scaling and dynamic voltage, which are being followed with performance benefits too.  Слайд 5  So, the term AI enhanced processors or simply AI processors encompasses both CPUs and discrete acceleration hardware, including GPUs, FPGAs, and purpose-built AI accelerators such as neural processing units (NPUs).  As we can see on given picture by AMD, these are the current biggest parts of modern computer processor.  Слайд 6  Let’s examine the benefits of having purpose-built AI accelerators.  The first one is - **Performance Enhancement**  Purpose-built AI accelerators in CPUs, such as Intel's Advanced Matrix Extensions (Intel AMX), significantly boost the performance of AI workloads. They were designed to speed up computations like matrix multiplications, which are crucial for deep learning and other AI tasks.  Second thing is - **Cost Efficiency**  CPUs with built-in AI accelerators can be more cost-effective compared to using discrete GPUs because they eliminate the need for additional hardware components, reducing the overall cost of ownership and the complexity of the system.  And last, but not least - **Energy Efficiency and Sustainability**  CPUs with integrated AI accelerators consume significantly less energy compared to GPUs, which is crucial for sustainability.  Слайд 7  Concluding that work we can point out the two main facts.  The integration of purpose-built AI accelerators within CPUs offers several significant benefits, particularly in the context of artificial intelligence (AI) workloads.  This undoubtedly leads to overall performance benefits since all processor loads can spread not only between the cores and threads of main processing units, but also these NPUs. | Слайд 1  Меня зовут Иван Иванов, Я учусь на кафедре Информационной безопасности МИЭТ и представляю доклад на тему «Процессоры с расширением Искусственного интеллекта».  Слайд 2  Цель данной работы – охватить / описать суть выделенного движка для Искусственного Интеллекта в современных компьютерных процессорах.  По ходу презентации мы будем обсуждать такие темы как закон Мура, ограничения в обработке, что же такое Процессоры с расширением ИИ и какая польза от специально сделанных ускорителей ИИ.  Слайд 3  Закон Мура гласит, что число транзисторов в интегральной схеме удваивается каждые 2 года.  Согласно данным компании Intel, собранным за последние 50 лет, они предсказывают, что темп роста числа транзисторов будет сохранятся точно так же, как это предсказывал закон Мура.  Но физически мы приближаемся к конечному размеру одиночного транзистора, так что может стать невозможной дальнейшая разработка более плотного расположения. Это приводит к факту, что мы могли бы использовать и другие методы и технологии для улучшения пределов обработки.  Слайд 4  Это приводит нас к очевидному вопросу: А какие есть другие способы нам или изготовителям дальше улучшать пределы обработки?  Есть несколько методов того, как изготовители пробовали и на самом деле улучшили скорости обработки.  Простейший из них – сделать сверхбольшие процессоры (больше привычных)  Вычислитель с набором упрощённых команд так же был реализован (так же стоит отметить ARM процессоры)  Многоядерность и многопоточность были спроектированы для того, чтобы улучшить в целом возможности и эффективность обработки.  Решение проблем с энергопотреблением и охлаждением могут привести к увеличению частоты и динамическому напряжению, что также приводит к повышенной производительности.  Слайд 5  Итак, термин Процессоры с расширением Искусственного интеллекта или просто процессоры с ИИ охватывают как ЦП, так и отдельное ускоряющее оборудование, в том числе ГП, ППВМ (Программируемая пользователем вентильная матрица) и специально сделанные ускорители ИИ, такие как Блоки Нейронной Обработки.  Как мы можем видеть на данном изображении от AMD, это нынешние крупные узлы/модули современного компьютерного процессора.  Слайд 6  Давайте изучим, какая польза от специально сделанного ускорителя ИИ.  Во-первых – **Улучшение производительности**  Специально сделанные ускорители ИИ в ЦП, такие как Усовершенствованное Матричное Расширение от Intel значительно улучшают производительность нагрузок на ИИ. Они были спроектированы для того, чтобы ускорить такие вычисления, как матричные умножения, которые критически важны для глубокого обучения ИИ и других его задач.  Второй пункт – **Эффективность стоимости**  ЦПроцессоры с специально сделанным ускорителем ИИ может быть более выгодным в сравнении со случаем использования отдельных ГП, поскольку они  Устраняют необходимость в дополнительном оборудовании, в целом уменьшая стоимость владения и сложность системы.  И последнее, но не менее важное – **Энергоэффективность и Надежность**  ЦП с встроенными ускорителями ИИ потребляют значительно меньше энергии в сравнении с ГП, что критически важно для надежности.  Слайд 7  Подытоживая работу, мы можем отметить два ключевых факта.  Добавление в ЦП специально сделанного ускорителя ИИ привносит несколько значительных преимуществ, в частности в рамках нагрузок на ИИ.  Это несомненно ведёт к улучшению производительности в целом, поскольку вся процессорная нагрузка распределяется не только по ядрам и потокам основного блока обработки, но и по этим блокам Нейронной Обработки. |