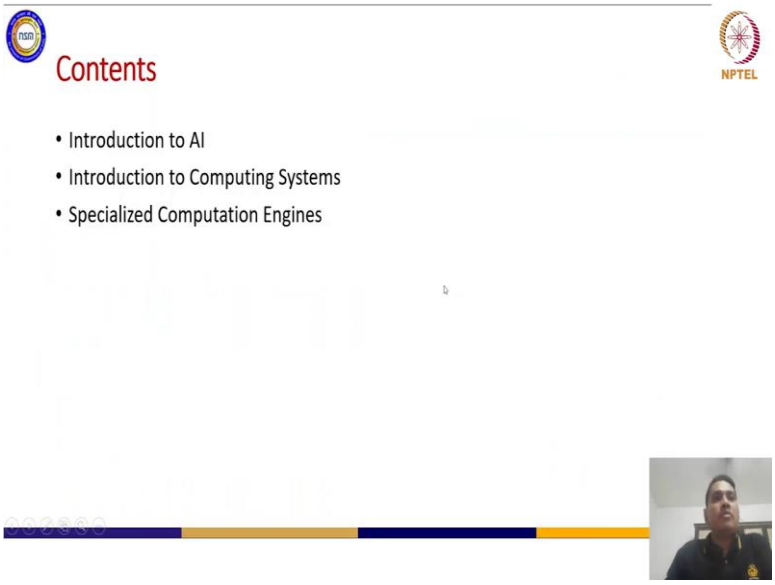
Applied Accelerated AI
Dr. Satyajit Das
Department of Computer Science and Engineering
Indian Institute of Technology, Palakkad

Lecture - 01
Introduction to AI Systems Hardware part 1

Hello and welcome to the Applied Accelerated Artificial Intelligence course, I am Satyajit Das. So, I will be handling the Introduction to AI Systems Hardware and system software. And I will be taking some of the SDKs with PyTorch and TensorFlow and yeah so that's about it and let's get started.

(Refer Slide Time: 00:40)

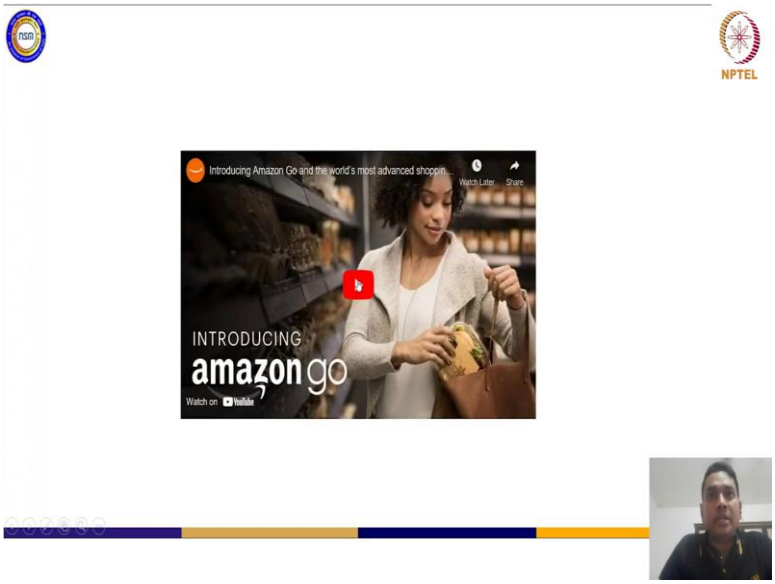


So, these are the contents of today's lecture. So, basically we will start with the introduction; we will see some applications of AI in modern days just to get motivated in the beginning and then we will talk about the computing systems from the perspective of AI.

So, from the perspective of running at artificial intelligence benchmarks, what are the modern systems available and how they can scale or how we can use them and what are the shortfalls that are there and how we can minimize the gap between the requirements and already available systems.

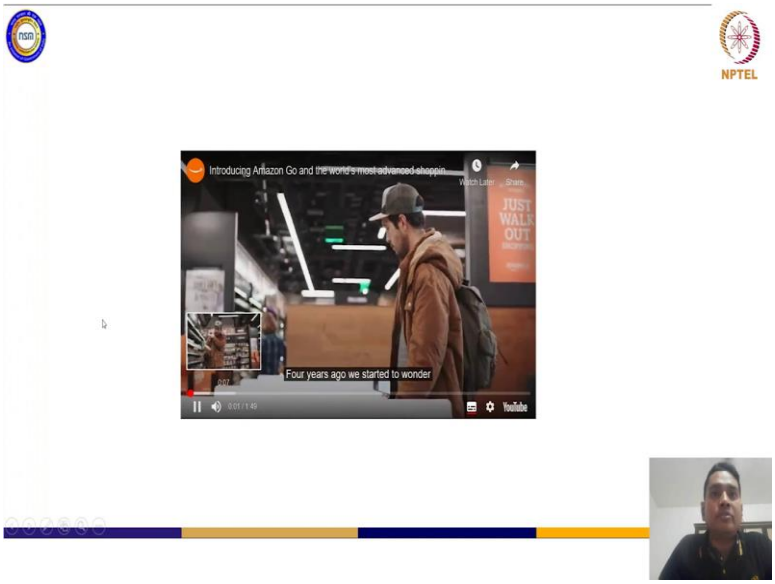
So, we will talk about those and of course, we will talk about some of the computing engines, we might not be covering all the computing engines that are available nowadays. But of course, we will try to cover some of these and to see of what are the things available nowadays.

(Refer Slide Time: 01:49)



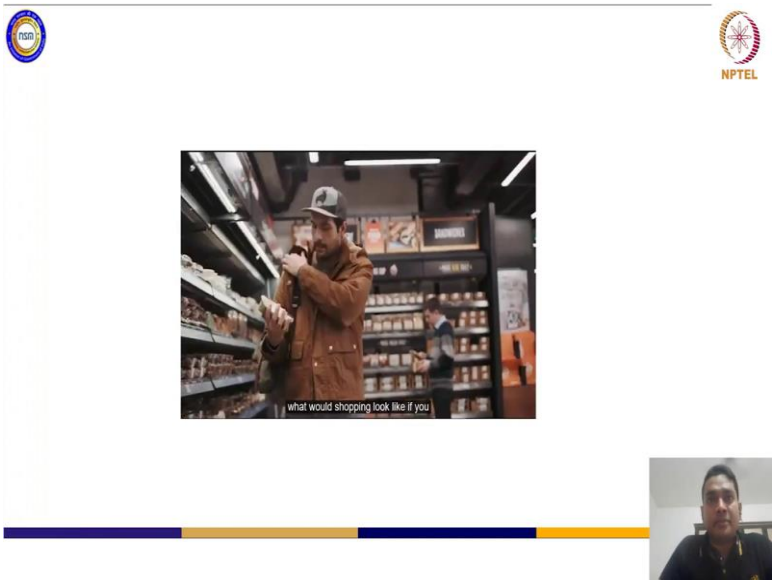
So, let's see one application as Amazon Go.

(Refer Slide Time: 01:56)

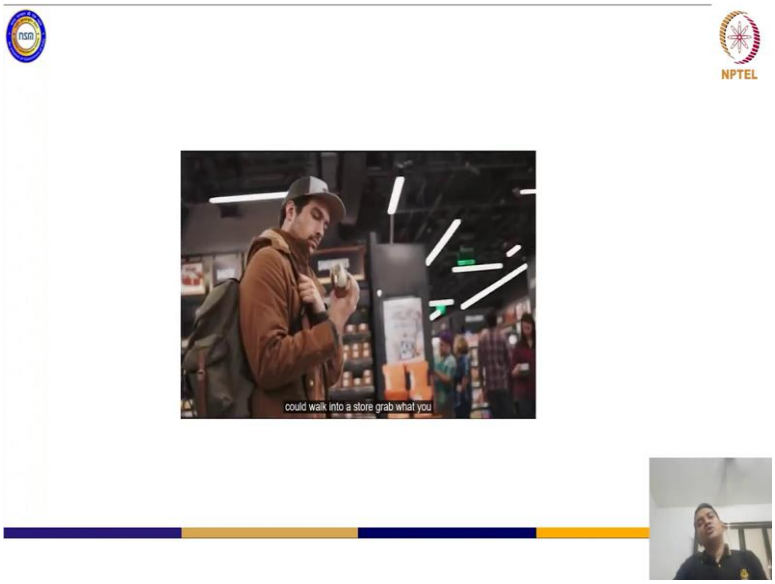


So, basically this Amazon Go provides you the flexibility to have seamless shopping experience.

(Refer Slide Time: 02:02)

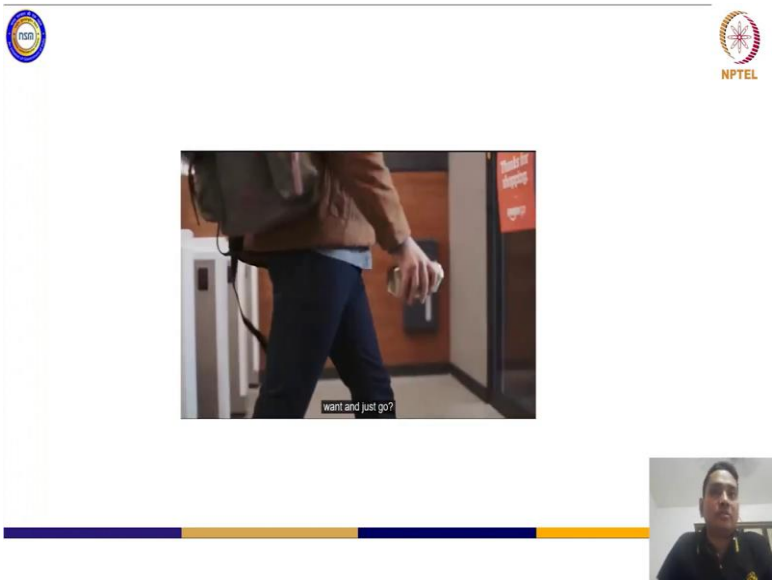


(Refer Slide Time: 02:04)



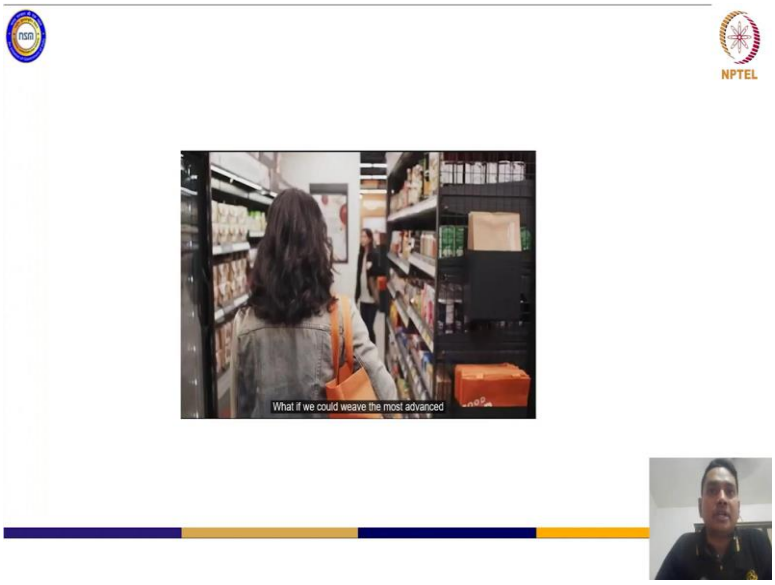
So, you go into one shop.

(Refer Slide Time: 02:06)



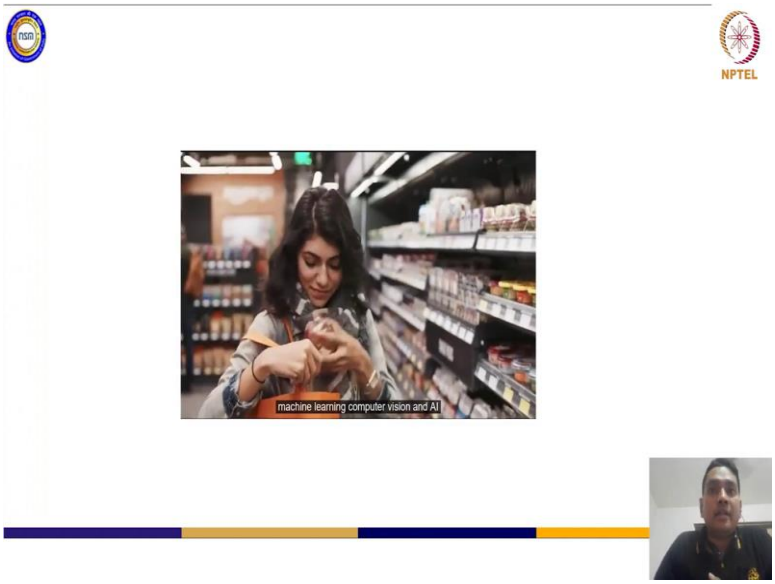
And you just take your things.

(Refer Slide Time: 02:08)

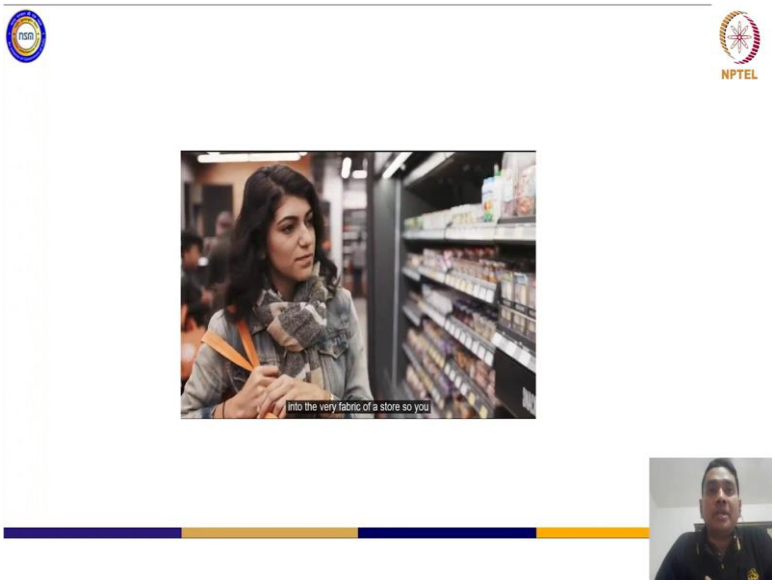


And you just get out of the shop.

(Refer Slide Time: 02:10)

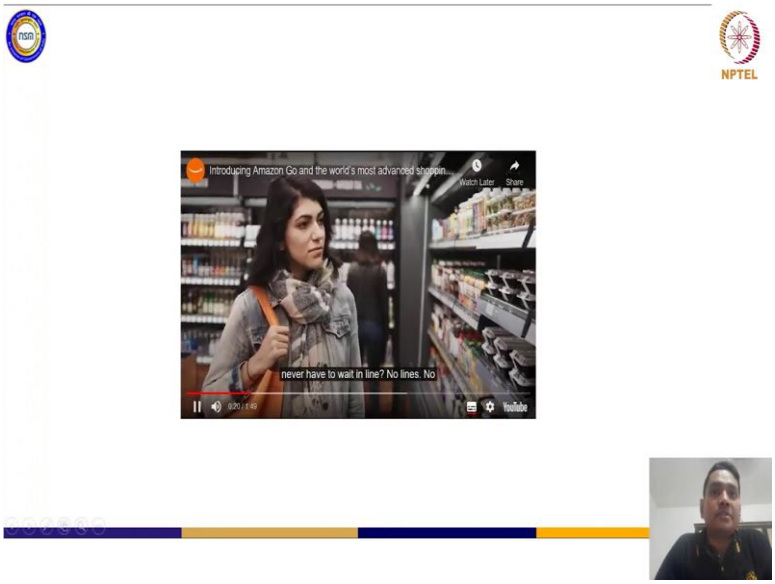


(Refer Slide Time: 02:13)

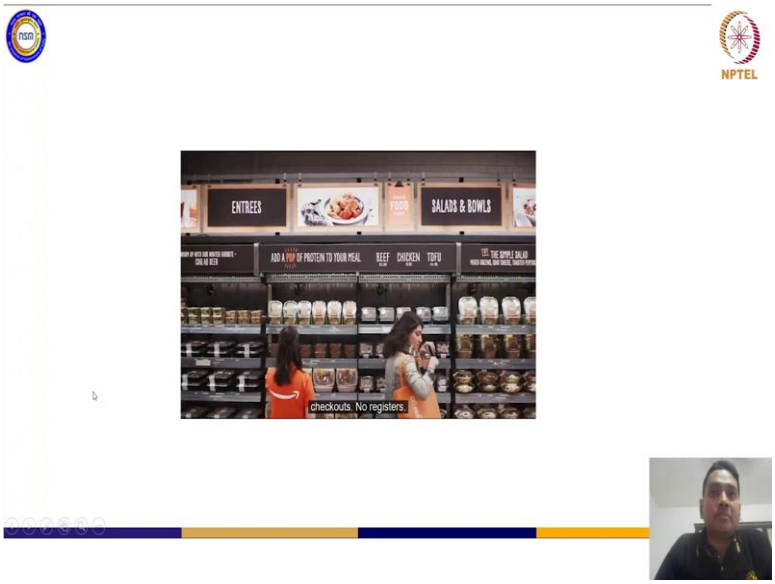


And your transaction will be automatically credited.

(Refer Slide Time: 02:15)

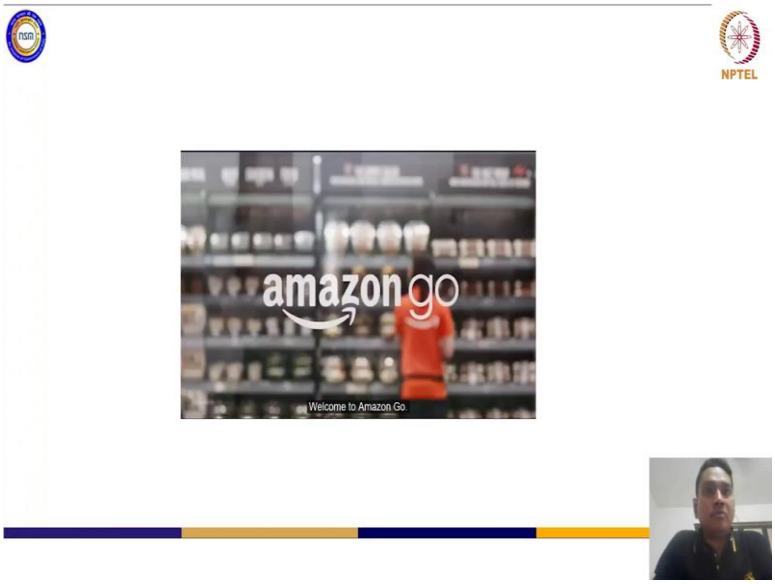


(Refer Slide Time: 02:19)

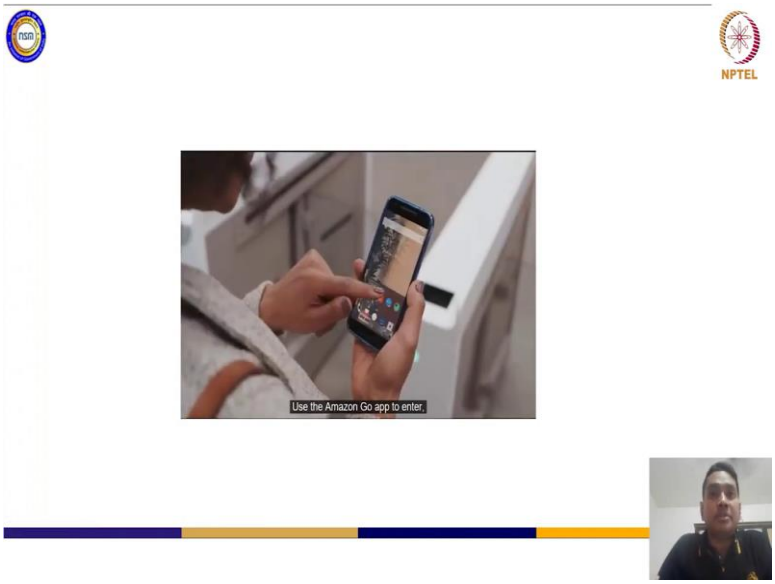


And you will have this seamless shopping experience lot of AI is applied here.

(Refer Slide Time: 02:23)

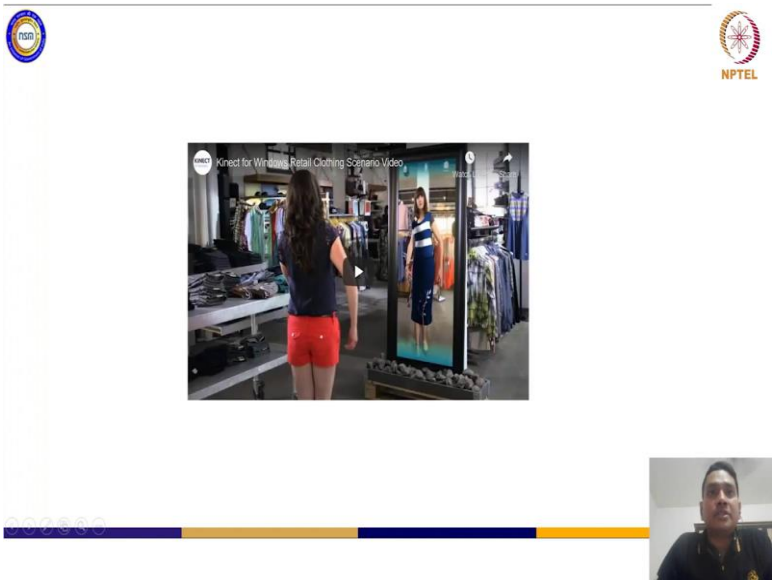


(Refer Slide Time: 02:27)

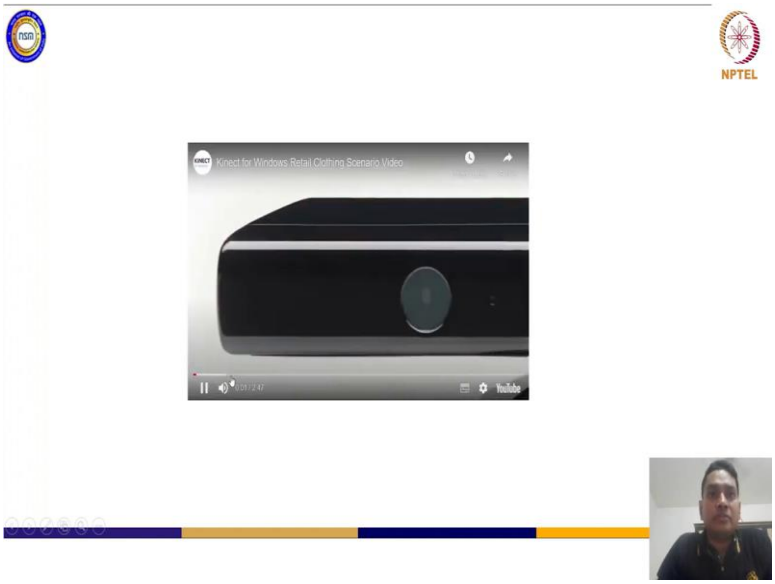


We will talk about what kind of algorithms or benchmarks that are being used.

(Refer Slide Time: 02:27)

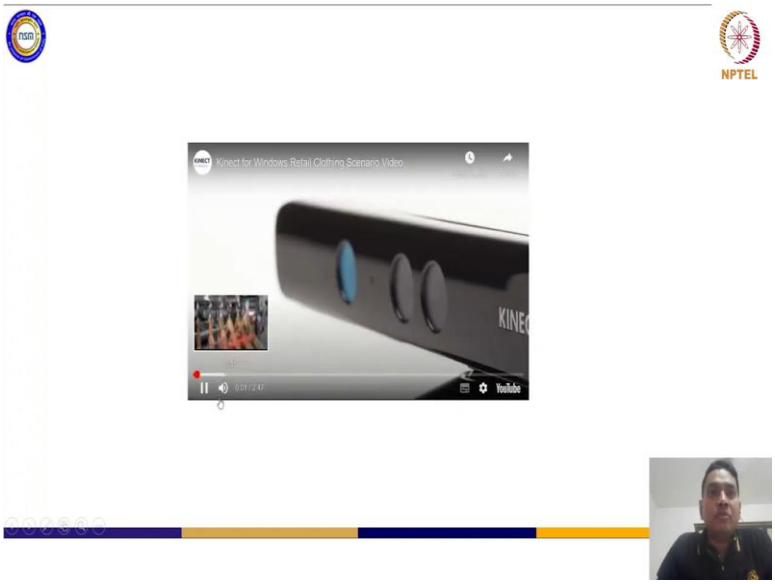


(Refer Slide Time: 02:32)

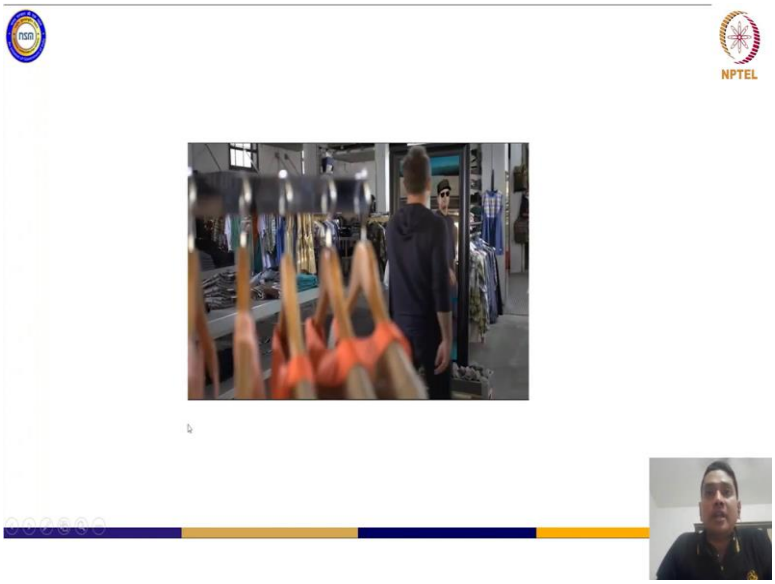


Another application is from this retail clothing scenario.

(Refer Slide Time: 02:34)

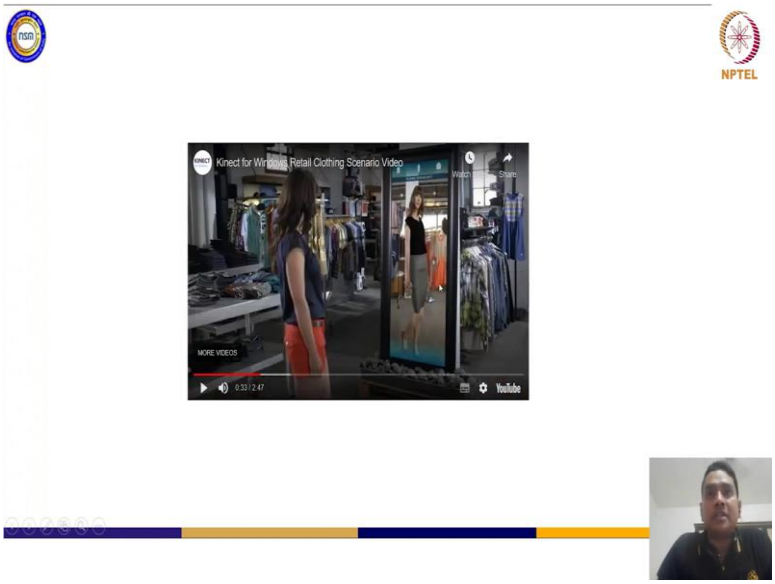


(Refer Slide Time: 02:34)



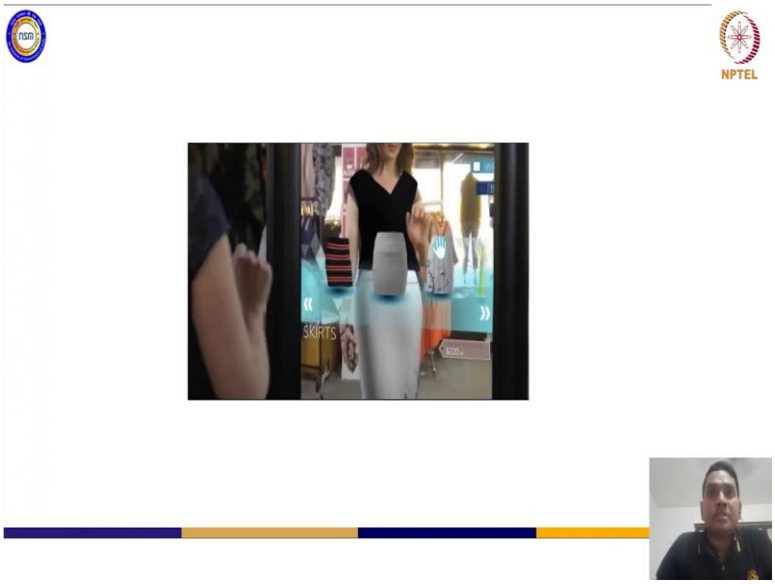
So, you go into one store and you without any contact you can just try out the clothings for your shapes and sizes that is available.

(Refer Slide Time: 02:41)



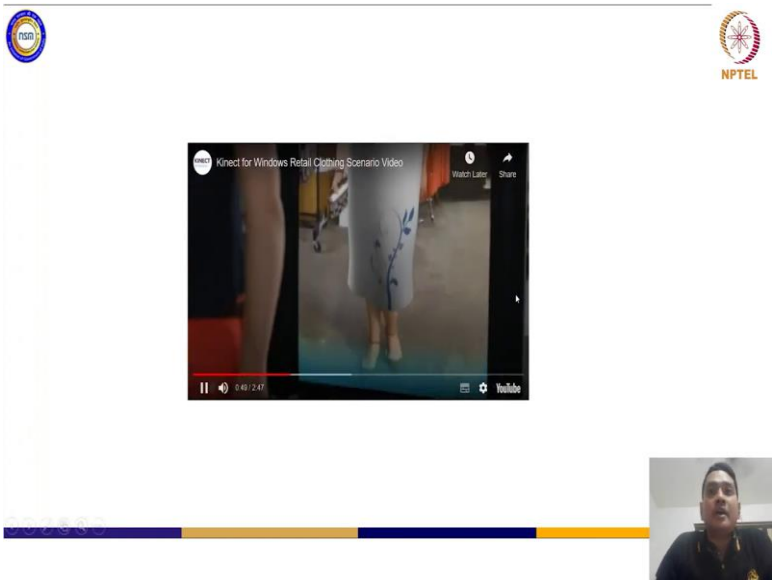
So, as you can see you can just stand in front of the mirror and you can try out several clothes depending on your requirements.

(Refer Slide Time: 02:55)

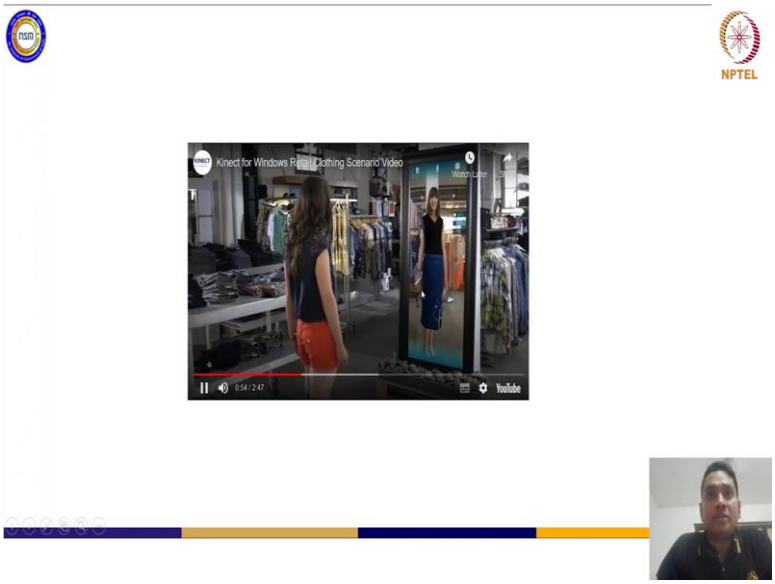


And like things and you can try it out.

(Refer Slide Time: 02:59)

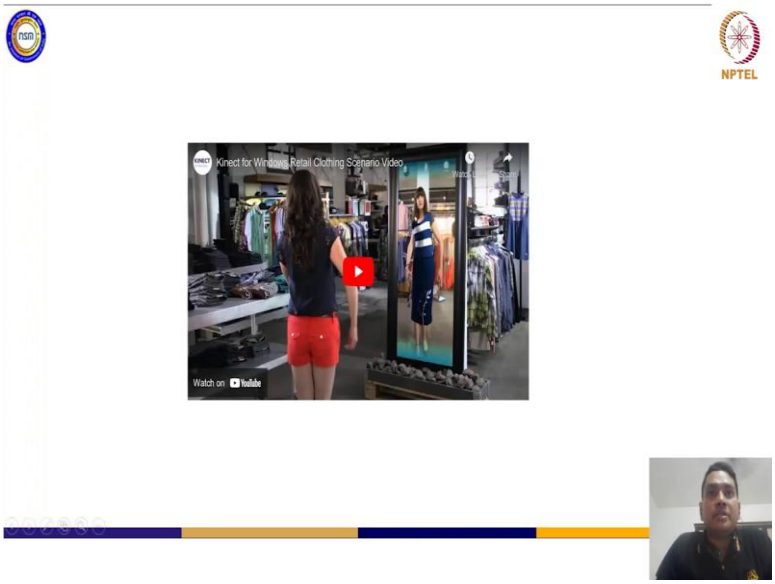


(Refer Slide Time: 03:03)

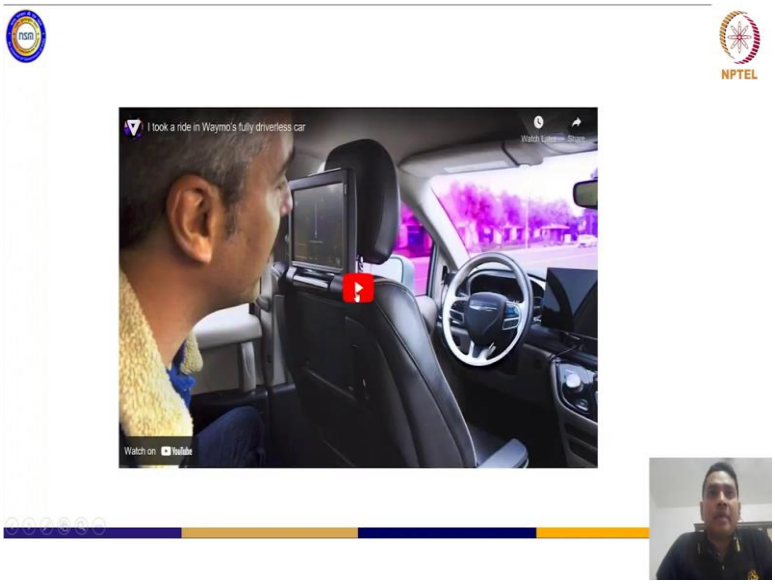


And you can have bought them as you go. So, without even going into the trial rooms you can have this seamless experience.

(Refer Slide Time: 03:15)

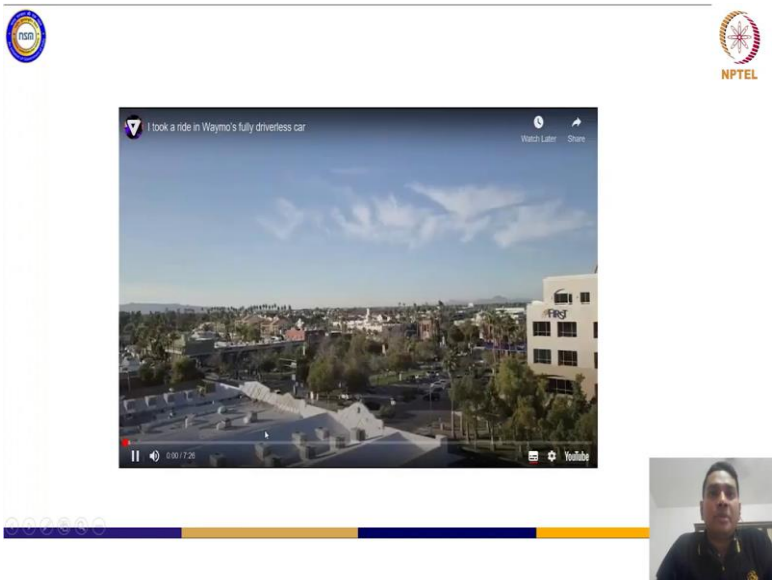


(Refer Slide Time: 03:17)

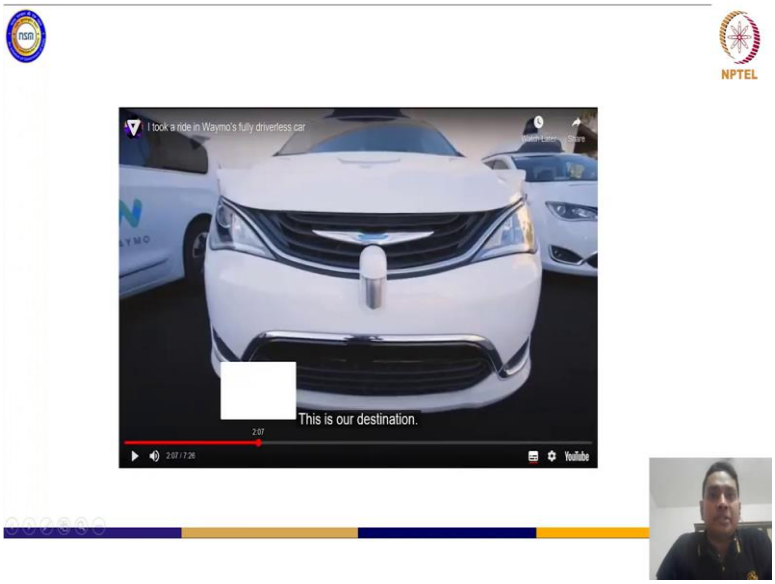


Again another application of course, is everybody knows nowadays the application of automated cars.

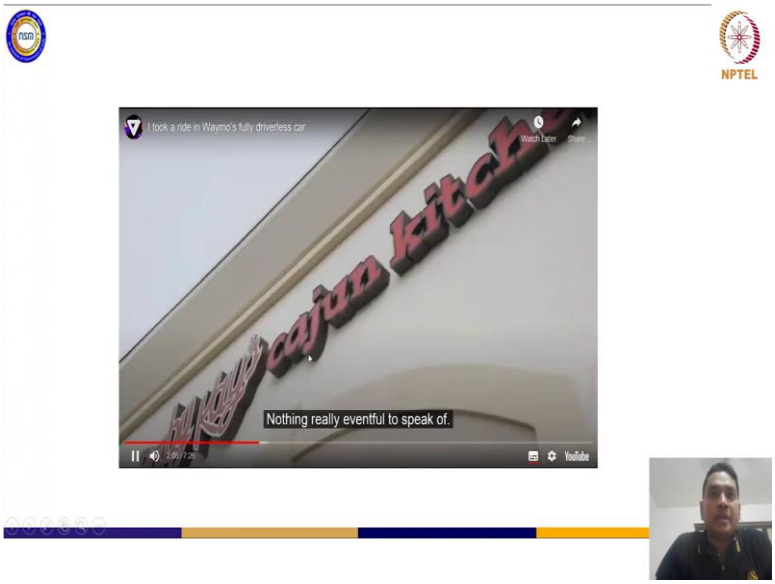
(Refer Slide Time: 03:24)



(Refer Slide Time: 03:25)

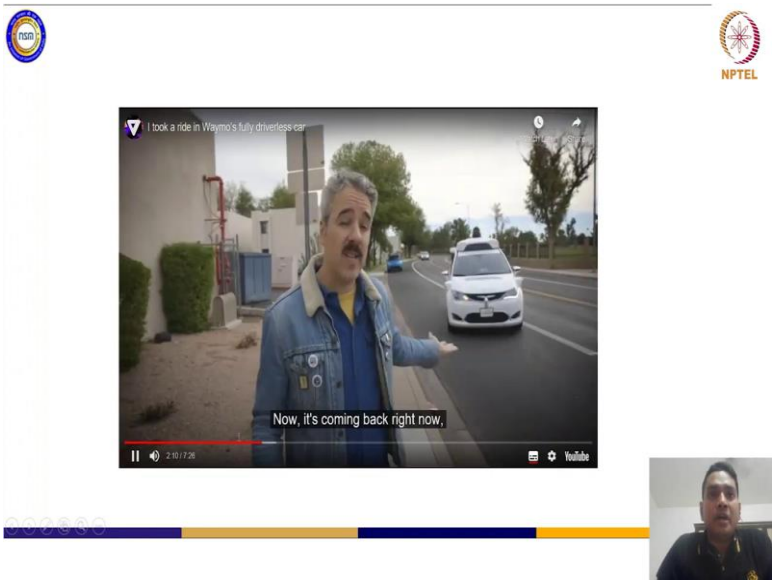


(Refer Slide Time: 03:27)

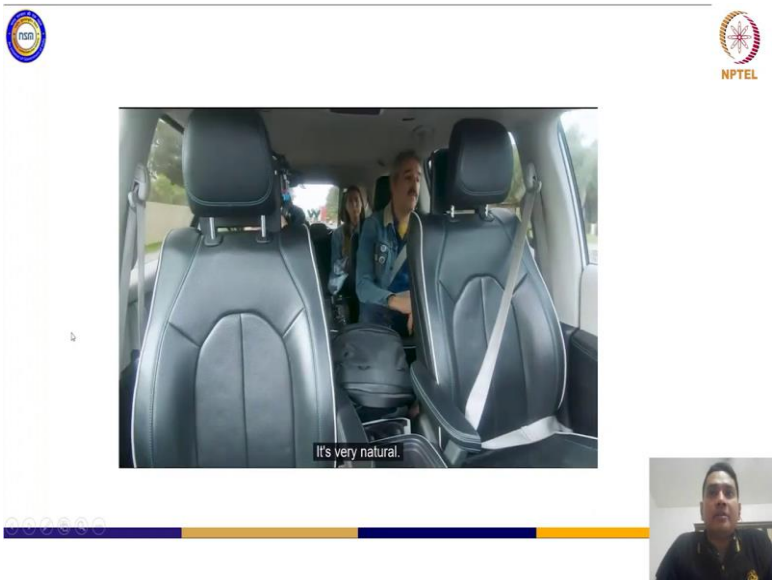


So, basically self driving vehicles.

(Refer Slide Time: 03:30)

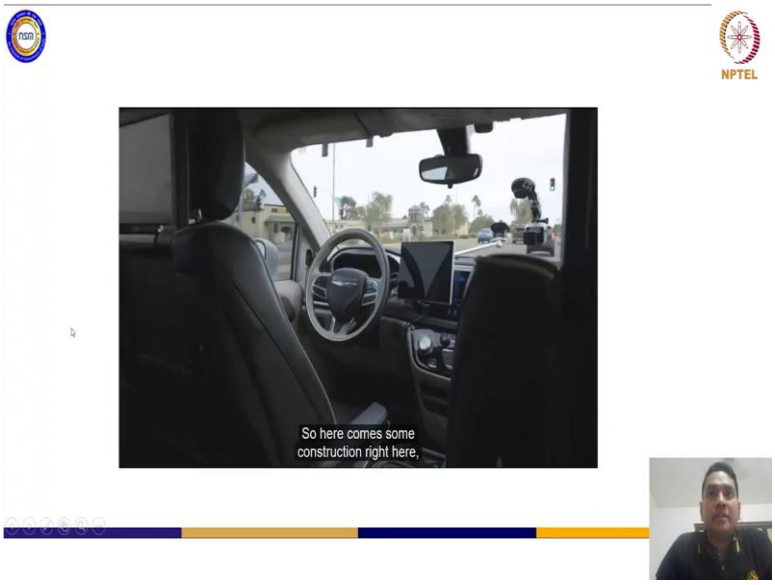


(Refer Slide Time: 03:31)

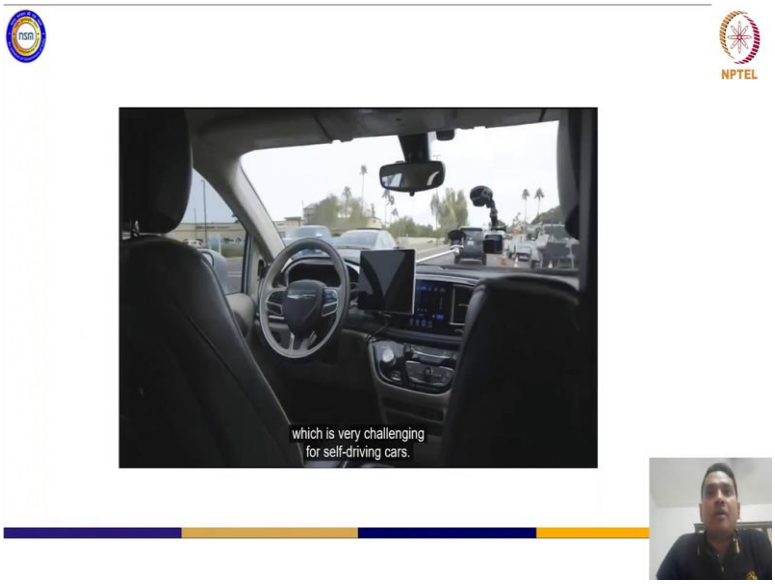


So, if you see different applications of computer vision and deep learning.

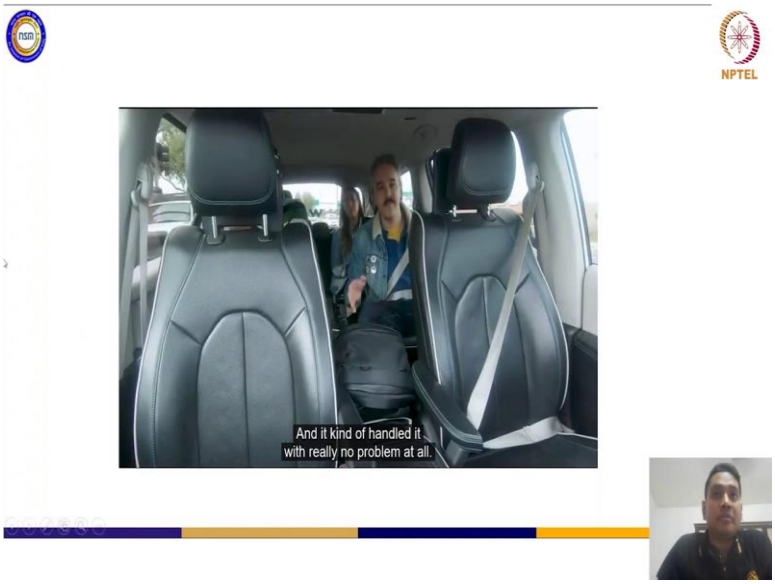
(Refer Slide Time: 03:37)



(Refer Slide Time: 03:40)

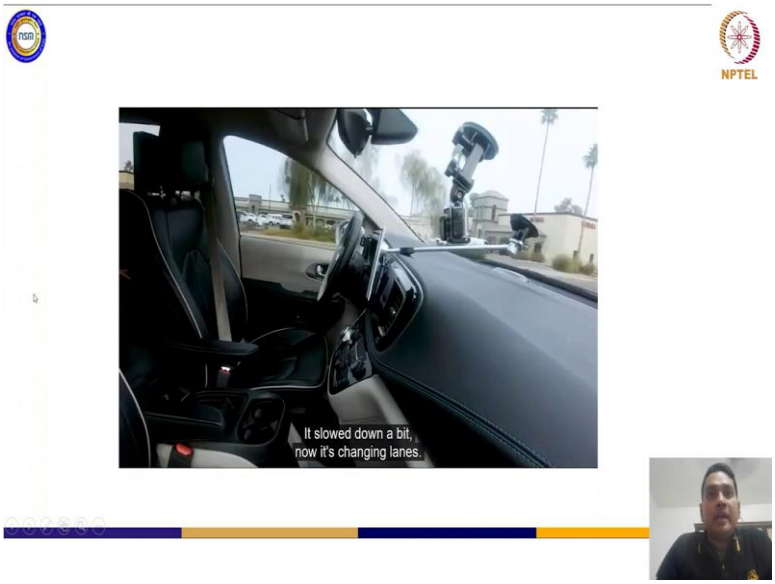


(Refer Slide Time: 03:43)

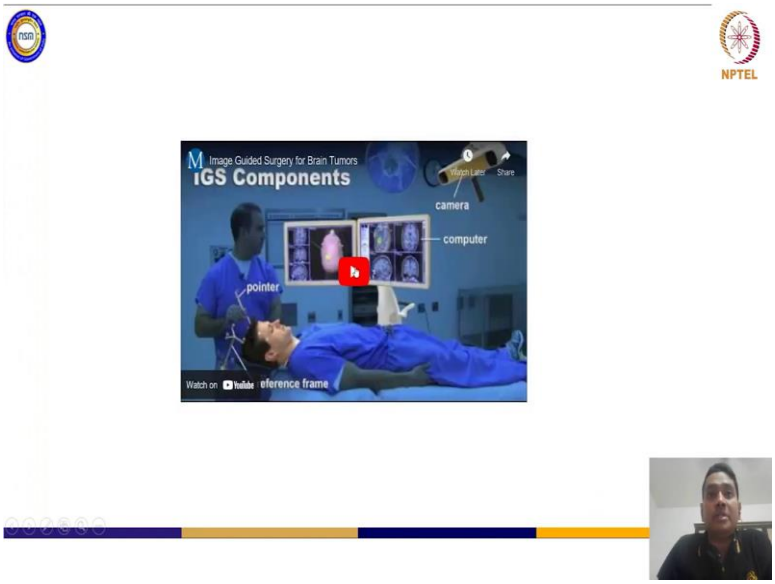


You will see it is most intense use of artificial intelligence in terms of application nowadays. Self driving cars.

(Refer Slide Time: 03:47)

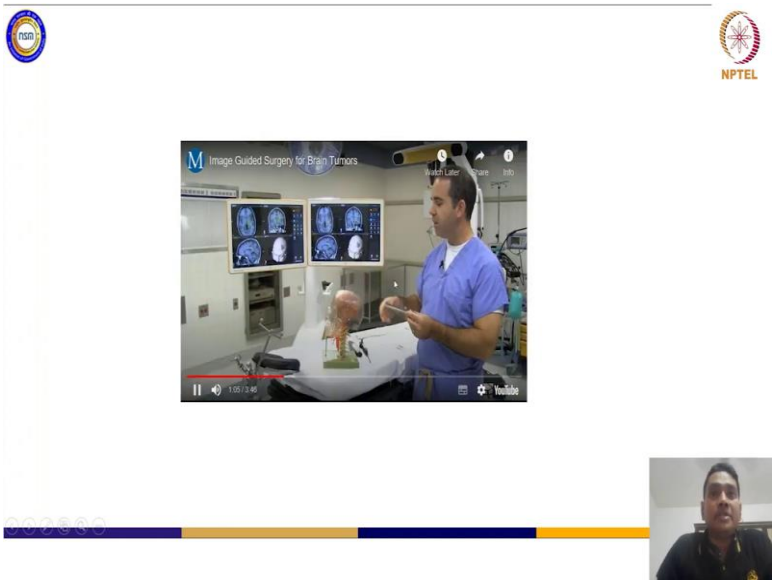


(Refer Slide Time: 03:48)

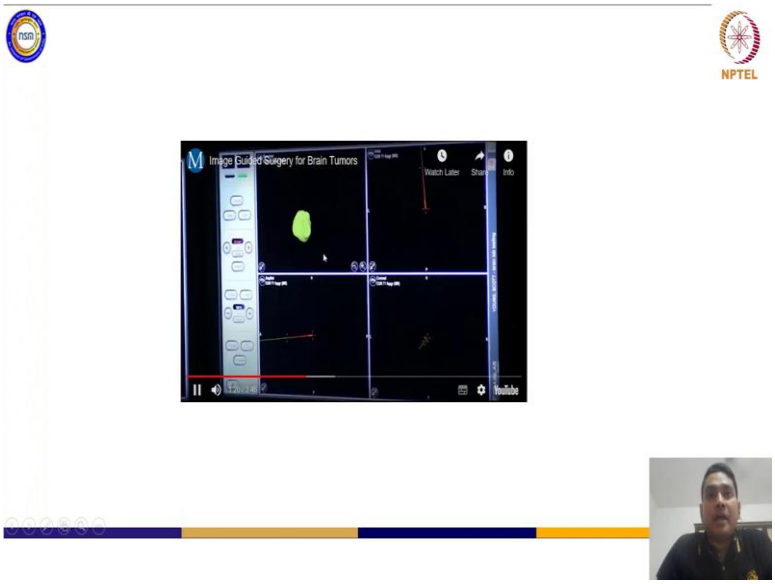


Again you have AI application in the area of medical imaging and automated operations.

(Refer Slide Time: 03:57)

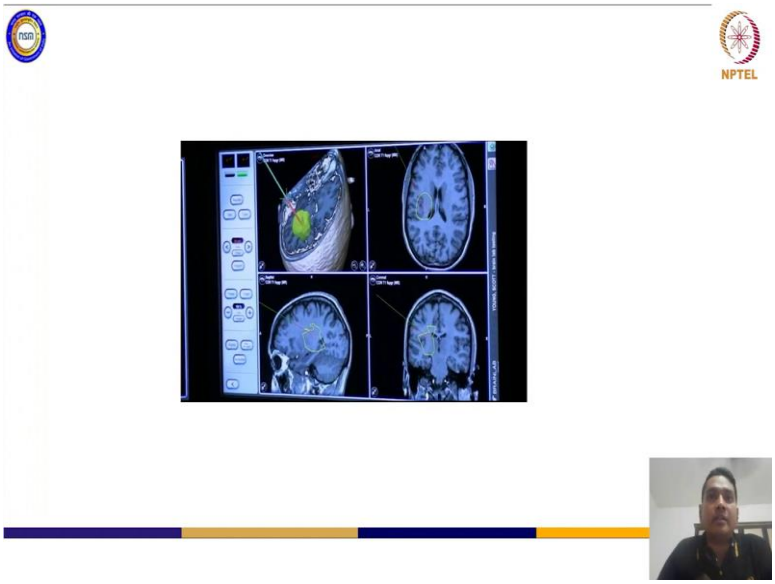


(Refer Slide Time: 04:01)



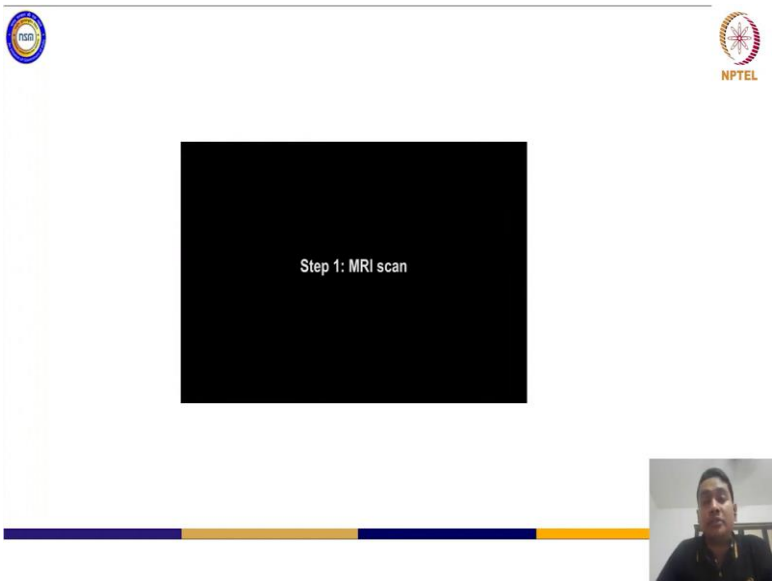
So, based on the images that are available you can actually have.

(Refer Slide Time: 04:03)



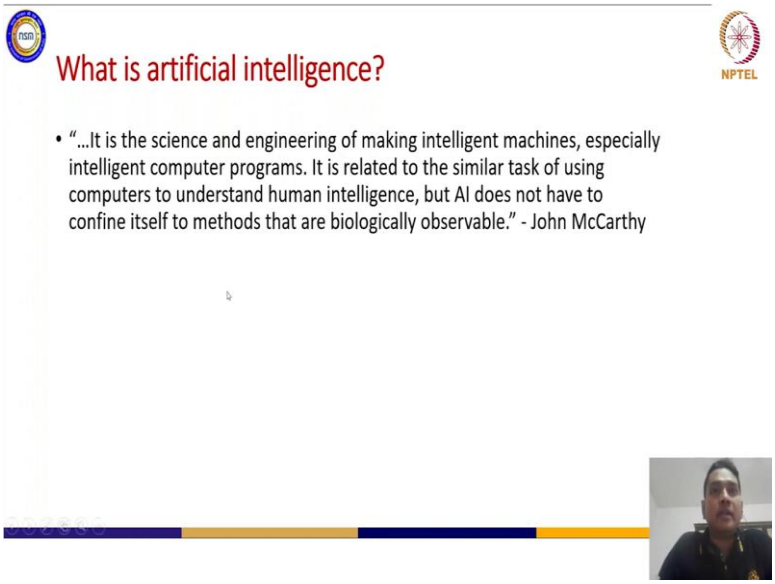
Like which way will be the most efficient and reliable way to go into one tumor that is there inside your brain.

(Refer Slide Time: 04:13)



And inside someone's brain and it can track that without any invasive measures right. So, these are the applications of AI and there is numerous applications you have heard of.

(Refer Slide Time: 04:29)

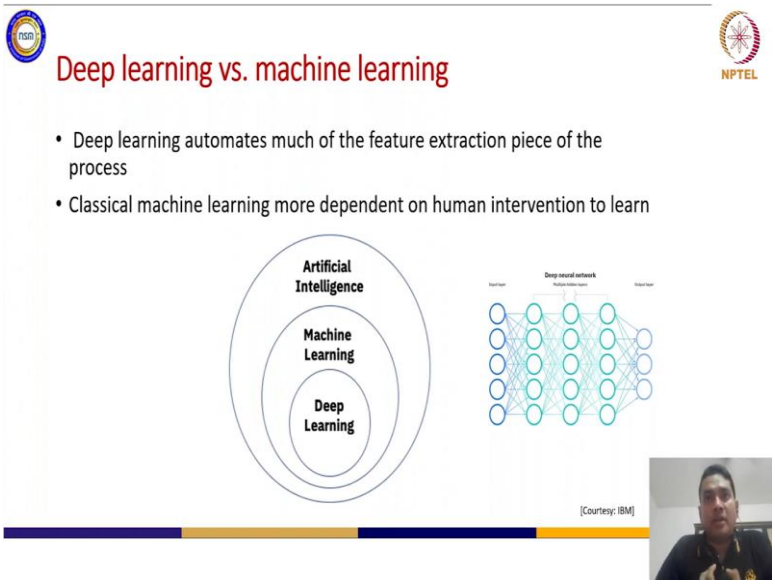


But the main question is what is artificial intelligence and here we have many definitions available. But mostly accepted definition was given by John McCarthy in the year of 2004 in one paper published from IBM and it says that it is the science and engineering of making intelligent machines especially intelligent computer programs.

So, the programs that are intelligent that are not anymore the rule based or conventional programming method right. And it is related to the similar task of using computers to understand human intelligence ok. So, that is the main purpose of emulating in human intelligence into the machines through these computer programs or intelligent computer programs. But AI does not have to confine itself to methods that are biologically observable of course, this is the generic acceptable definition of artificial intelligence.

But broadly it has evolved vastly from the earlier rule-based approach to solve intelligent problems to learning based approach. And these learning-based approaches became very popular with the advancement of new computing systems that are like GPUs and FPGAs and different systems like that.

(Refer Slide Time: 06:08)



So, as I was mentioning that learning based approaches to solve or to emulate human intelligence into computer programs have been very popular nowadays. And these are mostly widely named as deep learning and machine learning; of course, we will use them interchangeably and it is important to understand the nuances between ok. So, deep learning is basically comprising of neural networks which has more than three layers ok three or more layers.

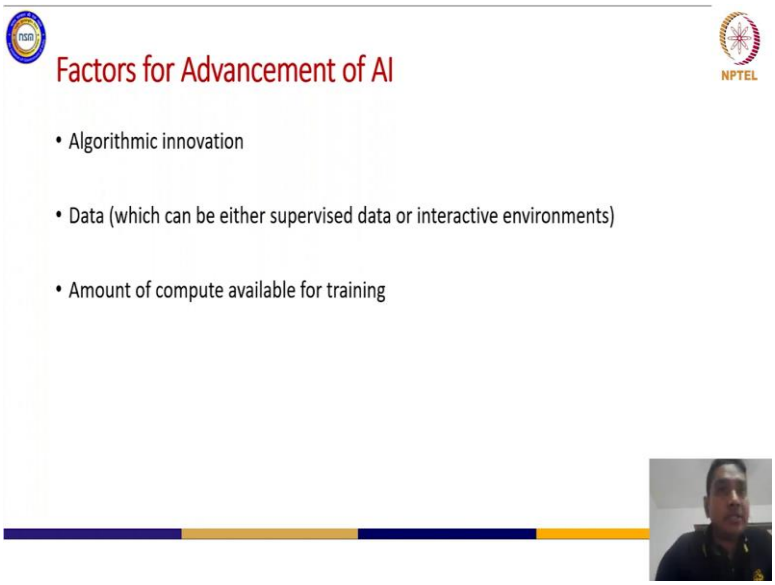
So, it includes the input and output layer as you can see here and deep learning automates this learning process by extracting the features from the data available. And that is completely or mostly automated because this does not need the human intervention as much as the classical machine learning needs that intervention of human dependence right.

So, that is kind of difference between these two concepts, but again deep learning is one subset of machine learning as you can see here right. And now the most important thing is that what kind of computations or what kind of complexity these benchmarks have. And if you understand that then it will be very easy to understand what kind of computing systems that we can engage them for right.

So, the configurations or if you do not understand the complexity of the algorithm, it will be hard to relate to the computing systems that are underlying or that will be running

your benchmarks right. So, from all these benchmarks that are available in deep learning that are mostly neural network based benchmarks.

(Refer Slide Time: 08:17)



But they are evolving day by day and that is mostly because of different algorithmic innovation. So, every year you can see hundreds thousands of paper being published from the algorithmic innovation point of view in machine learning and deep learning and their applications. Of course, when I say applications they are not confined to only deep learning or machine learning that they can be applied to let's say natural language processing or text processing, signal processing or maybe speech processing or got it right.

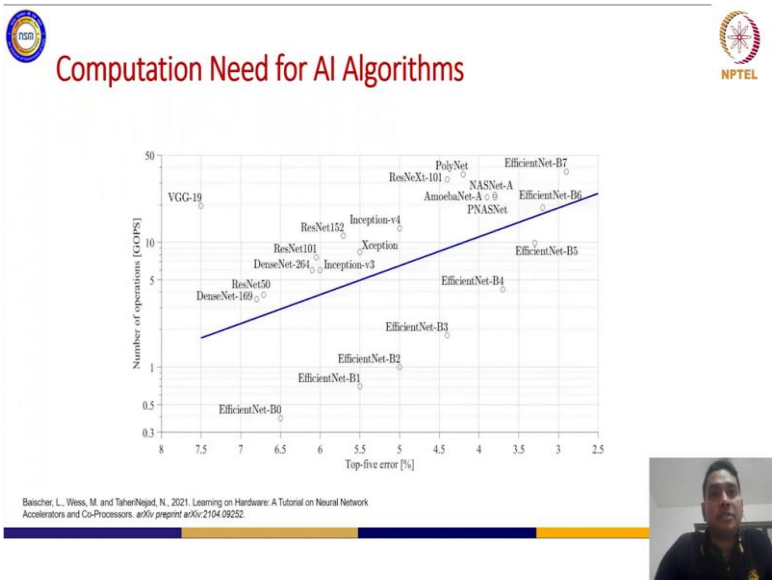
But of course, these core benchmarks will be used for different applications including the computer vision also that that application that we just saw as let us say in self driving cars or Amazon Go or different other applications. So, these algorithmic evaluation innovations are happening along with the more and more availability of data and nowadays we have abundance of data and in terms of learning the more data you have intuitively your learning will be better ok.

So, the amount of compute or computation you need to do you train these algorithms are also getting more or more; because you have more and more data and you have to compute or you have to process those data into the systems that is available at your

disposal, but of course, the amount increases and the compute intensity also will increase right.

So, now with this factor of advancement of AI, it is necessary to understand the computation complexity of AI algorithms right.

(Refer Slide Time: 10:18)



So, this paper was published last year in archive preprint as you can see here and of course, you will get to learn a lot about these neural network algorithms with basics and different algorithmic innovations as well as different systems that are available to process them this is a good read. So, in the reference I will give a link to this paper.

But if you see this graph, this graph shows a very like summarized way of representing these algorithms in terms of their computational complexity right. So, if you see the x axis, x axis in this graph represents what the top 5 error. Now what is top 5 error? Top 5 error is that these algorithms are supposed to learn something and depending on their learned parameters or whatever they have learnt for. So, basically the parameters that we call them or the model itself they will probably produce different possibilities for your output.

Let's say one classification problem ok. So, this algorithms that you are seeing here like EfficientNet, DenseNet, ResNet and all these different DNN or specifically CNN

networks or Convolutional Neural Networks they are supposed to provide you the classification of one data set called ImageNet ok.

Now what kind of possibilities they are producing whether that represents the true class or not. So, that is the measure of the error percentage and top 5 error is if your guess top five guess of these algorithms lie in that your true class of this particular image or video that you are trying to classify if they lie on their top 5 cases.

Then how much percentage of time they can guess that. So, that is the measure of this top 5 error. So; that means, if you want to increase the accuracy so, this is basically then the graph showing the accuracy versus the number of operations need to be processed for these benchmarks.

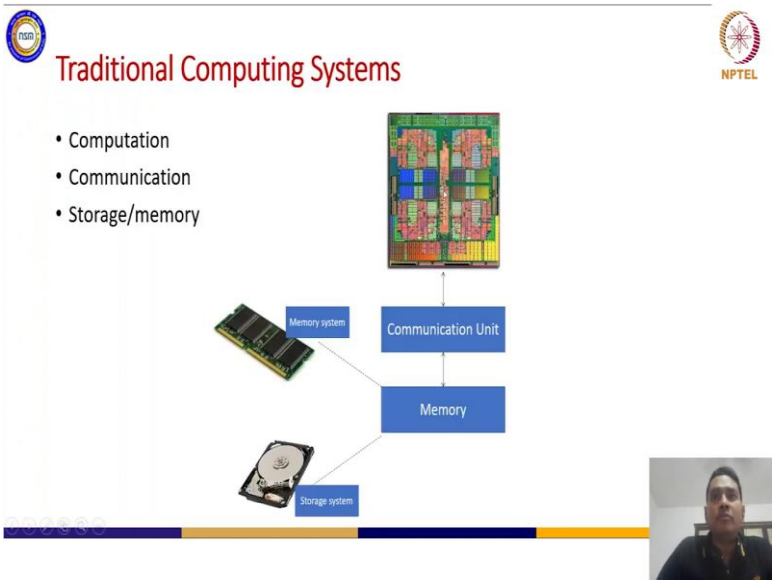
And these number of operations you can see in the y axis as represented as GOPS or giga operations per second you can see that in many places they are also referred to as gigaflops. So, floating point operations also the same as GOPS; because all the competitions will be in these benchmarks will be mostly for floating point arithmetic right.

So, now you have top 5 error you want to increase the accuracy of this benchmarks and you need to also increase the number of computations that you need to do for this benchmarks to gave more accurate. Now, this blue line you can see here is the accuracy measure right, it is linear.

So, if you want to get linear increase in the accuracy the computations you can see the evolvement of the computations get exponentially increased ok. And this is the factor you need to understand before going into implementation of these algorithms that you will see in the subsequent classes right.

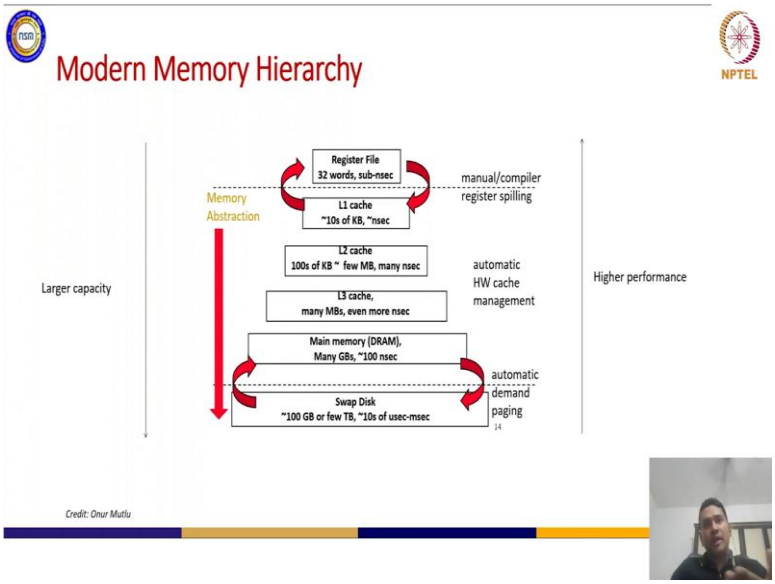
So, now this is one fact or one take away from this slide you will remember that to increase or to get linear accuracy you need to increase the computations exponentially right. Now we will look into the computing systems that are available and we will try to relate what kind of computing systems will be more suited for this kind of computational density or computational complexity that are being incurred by this AI benchmarks right.

(Refer Slide Time: 14:45)



So, of course, traditional computing systems as you know that we have computation engine and then we have memory, we have memory system different levels or hierarchy of memory is there, that we will see in the next slide and you have the communication unit. Now if your memory is on chip then you have on chip interconnect and if your memory is staying outside like dynamic memory like DRAM or your storage system. So, these are often housed outside of the chip and then you need off chip interconnect to get access to this memory right.

(Refer Slide Time: 15:24)



And if you see the hierarchy of the entire memory system in the modern computing system you have the register file at the lowest end. And then you have different levels of caches and L1, L2, L3 and then you have your main memory which is your RAM dynamic RAM. And then you have the swap disk to interchange programs between the swap disk and your main memory depending on the locality or availability of the program right.

And that is mostly automated and controlled by demand paging. So, most of these concepts also we will see in the subsequent class in system software, but the most important thing in this slide to look at that the speed of memory accesses. Now if you see in the register files the memory access speed is highest, then you have L1 cache in the range of nanoseconds.

Then you have many nanoseconds in L2 cache; then some more nano nanoseconds in L 3 cache and few hundred nanosecond in your main memory access. So, DRAM access and then you have 10s of microseconds to milliseconds in the access of swap memory or swap disk driven.

Now, in the modern systems of course, your swap disk is extendable to your remote storage through different gateways that you can increase and it can have several TBs even petabytes of swap disks available. But just to look at the local system then that you can have the sizes vary from words to KBs to several MBs in the last level of cache and then several GBs in your dynamic memory or dynamic RAM right. Now why this hierarchy is necessary?

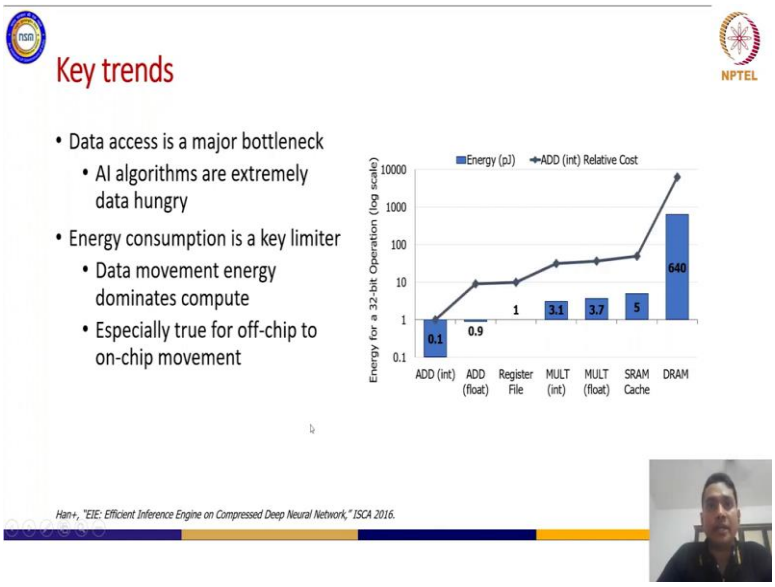
Because if you see the L1 level of cache this gives you a very faster access to the computing unit because the computations are much more faster. And you need the data available to the computation unit much more faster than the dynamic RAM; because dynamic RAM you can see the access is higher right the access time is higher.

So, at the lower level you go the access time will increase the performance of access time will increase; that means, you will get access in less time. And if you go in the upper level the performance will degrade and also you can see that it can house larger size ok. So, the sizes also you can see.

Now, why this is important? This is important because the computation unit now see that it has now that illusion of high or several level of memory is giving you like high volume ok. So, larger space so, this illusion of larger space with very short access time is giving you this hierarchy ok. So, this is why this hierarchy is important now; why it is important from AI computational AI benchmarks point of view? Because from AI benchmarks we have seen that we need lot of data to be accessed for training these benchmarks right.

And when we talk about data, memory is the first thing it will come into mind right because our memory is where you will store the data. So, what modern systems are doing is that, because if you see that access times will be very low for this L1 level of cache then you keep it the closest to your computing in it that that, but your processor of code. And then you keep your L2 cache then you keep L 3 cache and you can keep your main memory or DRAM and swap disk off chip ok.

(Refer Slide Time: 19:51)



And that is what actually happening nowadays and as I was mentioning that the benchmarks are bottlenecked by the data access the AI algorithms are extremely data hungry, if you can see that how much data computation it needs per second right you have seen the graph before. So, for these memory accesses another key factor is the energy consumption.

So, size, the access time that we have seen in the previous slide now from the point of view of energy consumption this memory hierarchy is also important; because the more

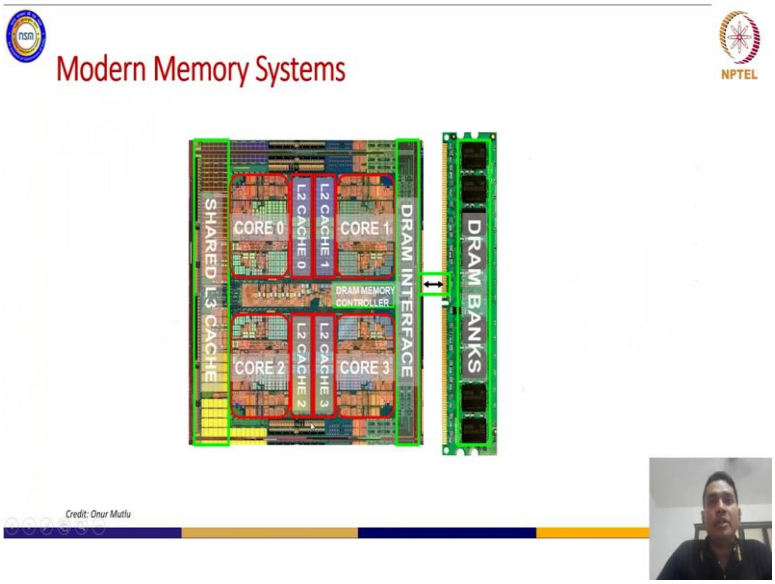
energy you will spend or accessing memory the more heat you will generate and you need more cooling system to employ or deploy to cool your systems and it is much more cost intensive than generating one system or chip right.

So, that is why energy consumption is a key limit of this data movement and data movement is itself is energy hungry; because if you see this graph of the data access energy consumption versus data computation. So, basically this in x axis you will see that these are the different operations. So, of course, AI benchmarks or DNN benchmarks are widely dominated by addition multiplication and multiply accumulates. So, addition and multiplication operations energy consumption you can see here.

And DRAM access and SRAM access so, SRAM is static RAM which is which technology is mostly employed in the caches. And for rams your drams you can see DRAM access and SRAM access the energy consumption is much higher. And if you go into DRAM it is even several orders of magnitude higher than the computation itself. And if you just compare with add operation this DRAM axis needs almost 6400 times of energy to access this DRAM right.

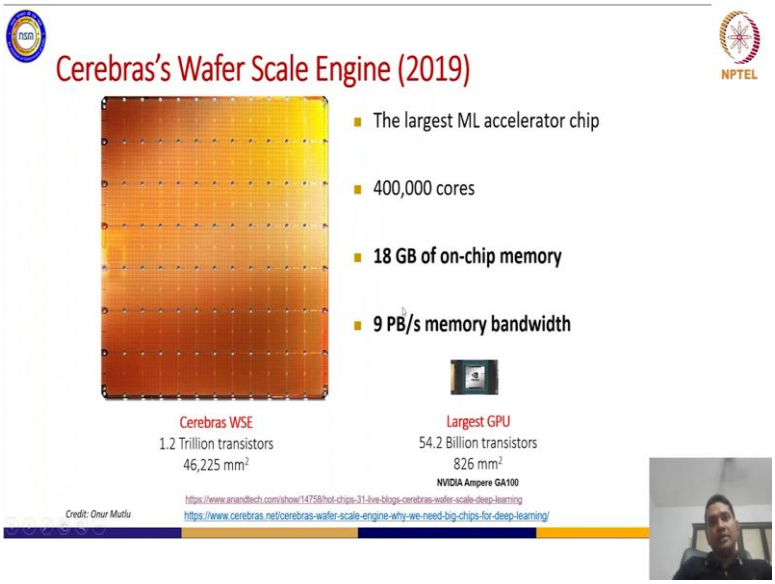
So, SRAMs or cache based memory that are on chip because they give you very short access time that we have seen in the previous slide. And now you can see here SRAM is also giving you much more energy efficiency in terms of memory accesses and DRAM is giving you much more energy consumption right. Now what means this is the trend that you can see here, but what this to be done.

(Refer Slide Time: 22:57)



So, modern memory systems what they do is that. So, they employ the L2 caches the L1 caches inside the core itself. So, in this cores that you are seeing here core 0 core 1 core 2 core 3. So, this is the core that is available and inside this cores you have L1 cache it is not particularly visible here, but L2 cache you can see prominently L2 cache level cache you can see on chip and then it has shared L3 level of cache on chip and DRAM is a kind of object because you need higher memory density and it needs to deploy much larger space of memory ok.

(Refer Slide Time: 23:46)



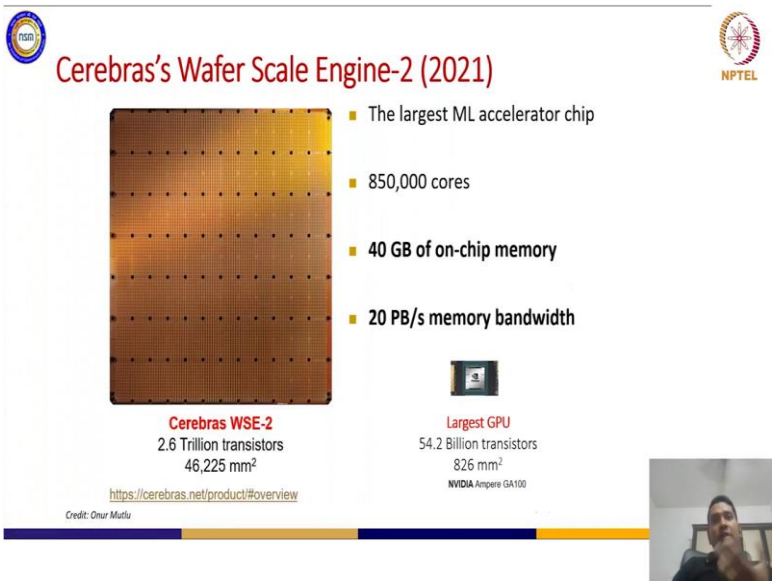
Now, if you see from the AI computation point of view, I will just give you one example of how much memory that are deployed inside the chip itself to get a larger memory bandwidth for running the AI benchmarks. So, this chip or this proposal was published in the year of 2019 and this is Cerebras Wafer Scale Engine.

So, basically this whole chip is basically one wafer itself and all the other chips that you see are basically separate; in one wafer you will have several or several number of chips, but this is a wafer scale engine and this is one ML compute engine. Basically it has a lot of multiply and accumulate unit stored as or arranged as an array of processing engines.

And this particular engine has 400,000 of this kind of cores ok and on chip memory it has around 18 GB of on chip memory. And this if you compare the size of this chip compared to the GPU that are available. So, the largest GPU that is available nowadays is NVIDIAs Ampere GA100 GPU; and that has around 5 54.2 billion transistors. And this particular WSE or Cerebras WSE is having 1.2 trillion of transistors.

So, you can imagine that how much bigger it is in size. And as well as to accelerate to accommodate all the computations that needed for different AI benchmarks that you have seen before if it employs 18 GB of on chip memory and that gives it 9 petabytes per second memory bandwidth. So, 9 petabytes per second data that you can access of course, this is a full precision data that we are talking about.

(Refer Slide Time: 25:55)



Now, one next generation of this wafer scale engine was published in 2021 and that has 850000 cores and that can deploy 40 GB of on chip memory and having a 20 petabytes per second of memory bandwidth. So, you can imagine how much the memory it is employed ok. So, just to have this shorter access time and reduced energy and. So, all these things we have seen in the previous slide right.

But of course, these are very highly specific of ML accelerators or ML compute engine or AI engine; must we will see much more generalized systems that are available and mainstream devices that are available ok. So, that brings us to the section, where we will talk about this specialized computation engines.