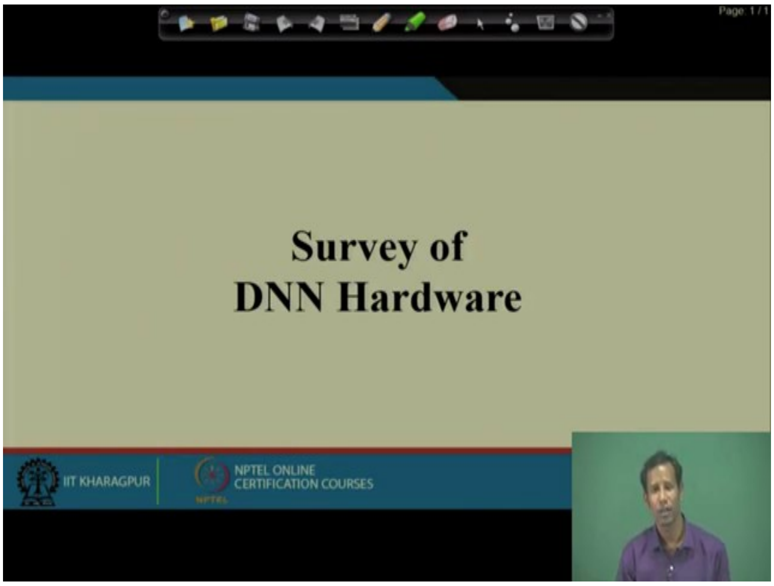
Architectural Design of Digital Integrated Circuits
Prof. Indranil Hatai
School of VLSI Technology
Indian Institute of Engineering Science and Technology, Shibpur, Howrah

Lecture – 67
Hardware for Machine Learning: Design Considerations Design Tips (Contd.)

Hello, everyone welcome to the course on Architectural Design of ICS. So, in the last class we are discussing about the Hardware Design for Machine Learning or artificial intelligence. So, there are wide applications of artificial intelligence in various fields. So, that is why we need to developed one efficient algorithm first and then we need also efficient hardware which will be designed so that we can make the applicability of artificial intelligence to everywhere. So, the overview of deep neural network actually we have already seen.

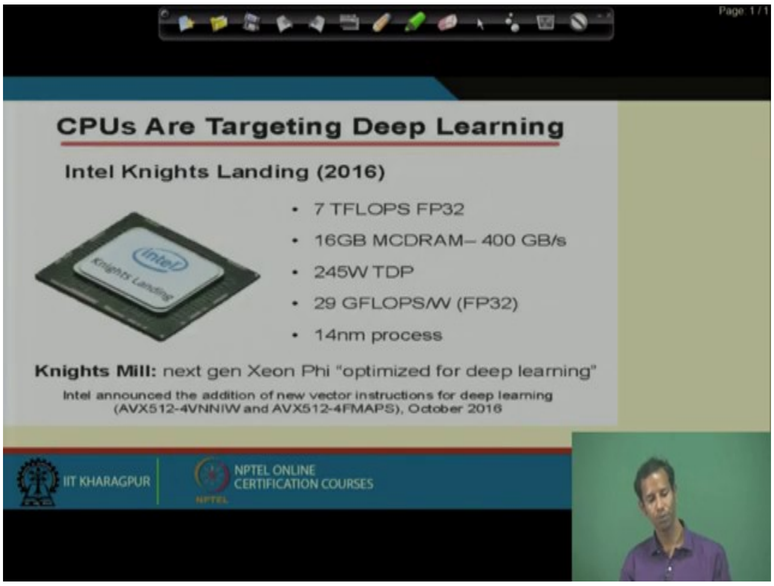
(Refer Slide Time: 01:15)



Survey of
DNN Hardware

So, the second thing is that now what are there? Actually if we consider that how many deep neural network hardware till now the people have already developed or people are still trying to develop?

(Refer Slide Time: 01:35)



CPUs Are Targeting Deep Learning

Intel Knights Landing (2016)

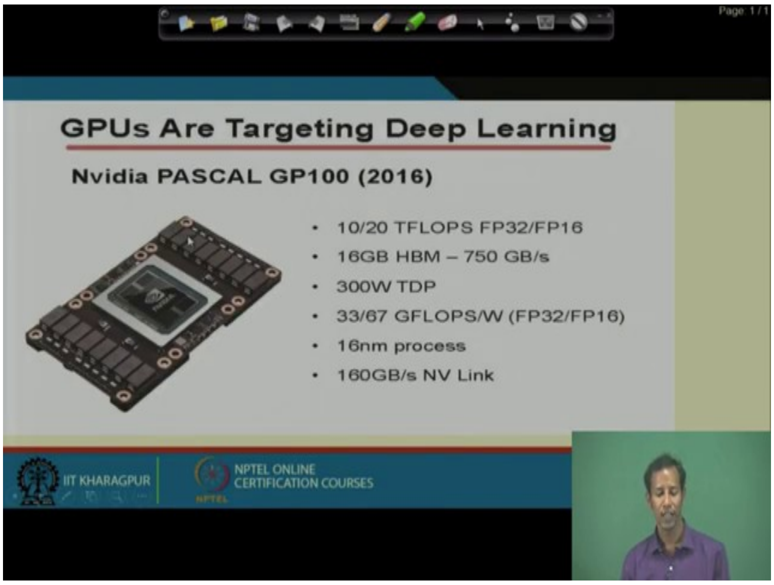
- 7 TFLOPS FP32
- 16GB MCDRAM– 400 GB/s
- 245W TDP
- 29 GFLOPS/W (FP32)
- 14nm process

Knights Mill: next gen Xeon Phi "optimized for deep learning"

Intel announced the addition of new vector instructions for deep learning (AVX512-4VNNIW and AVX512-4FMAPS), October 2016

So, if you see that this Intel has in 2016, they have developed one processor they came up with deep learning processor which is this Knights Landing. It has this 7 teraflops inside of this processors which is having this 16 GB MCDRAM and then it is being made in 14 nanometer process ok. So, this is the Knights Mill is the next generation Xeon Phi optimized for deep learning, so that means, process.

(Refer Slide Time: 02:13)



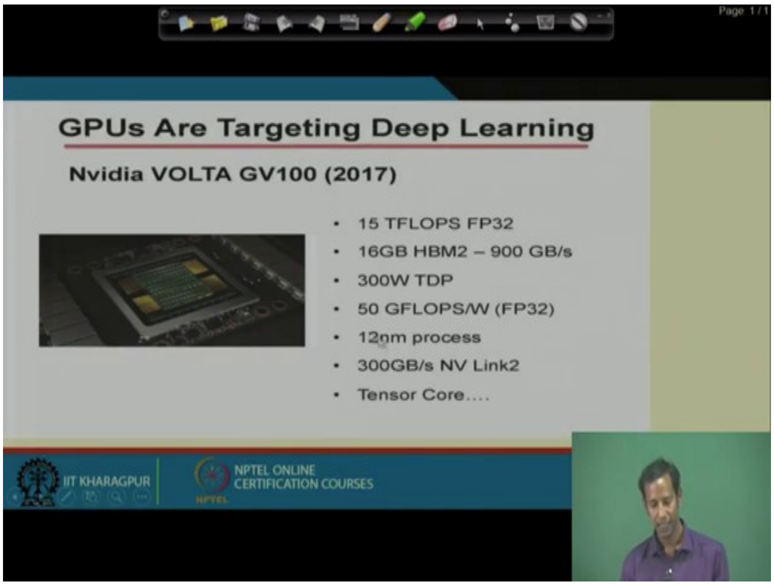
GPUs Are Targeting Deep Learning

Nvidia PASCAL GP100 (2016)

- 10/20 TFLOPS FP32/FP16
- 16GB HBM – 750 GB/s
- 300W TDP
- 33/67 GFLOPS/W (FP32/FP16)
- 16nm process
- 160GB/s NV Link

So, then we are having this Nvidia PASCAL which is another example of Deep Learning processor, it is also made by 16 nanometer process.

(Refer Slide Time: 02:24)



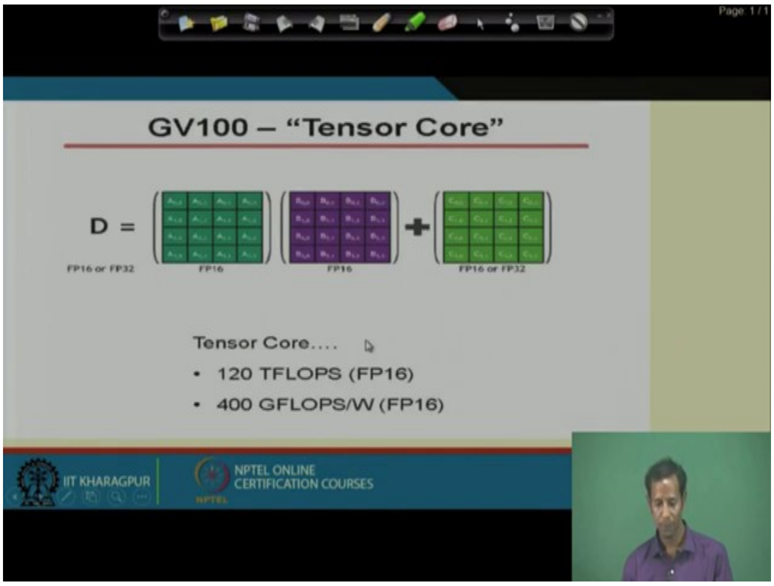
GPUs Are Targeting Deep Learning

Nvidia VOLTA GV100 (2017)

- 15 TFLOPS FP32
- 16GB HBM2 – 900 GB/s
- 300W TDP
- 50 GFLOPS/W (FP32)
- 12nm process
- 300GB/s NV Link2
- Tensor Core....

Then we are having this Nvidia VOLTA which also has the capability to to do the Deep Learning and it is being made by 12 nanometer process technology node.

(Refer Slide Time: 02:36)



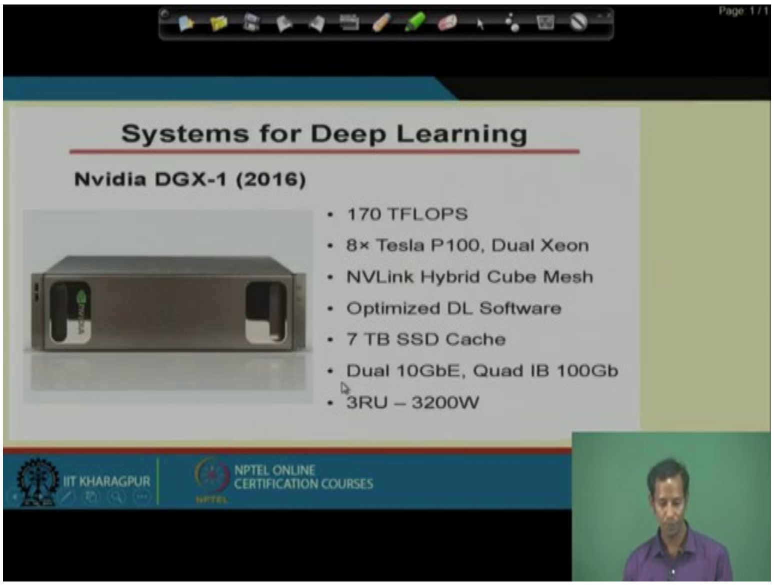
GV100 – "Tensor Core"

Tensor Core....

- 120 TFLOPS (FP16)
- 400 GFLOPS/W (FP16)

Then actually, we are having this GV 100 Tensor Core which is having this capability of 120 teraflops, operation it can do.

(Refer Slide Time: 02:48)



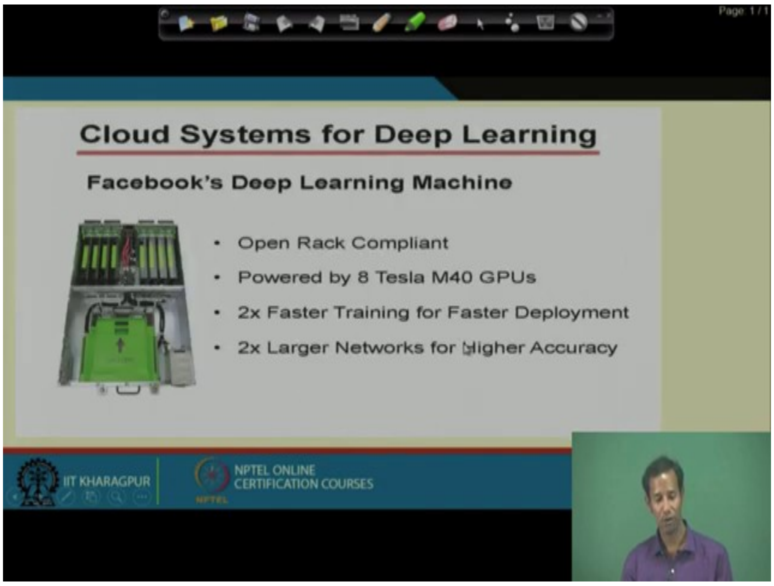
Systems for Deep Learning

Nvidia DGX-1 (2016)

- 170 TFLOPS
- 8× Tesla P100, Dual Xeon
- NVLink Hybrid Cube Mesh
- Optimized DL Software
- 7 TB SSD Cache
- Dual 10GbE, Quad IB 100Gb
- 3RU – 3200W

Then we are having this Nvidia DGX-1 in developed in 2016 and it is basically having 170 teraflops to do the corresponding operation.

(Refer Slide Time: 03:05)



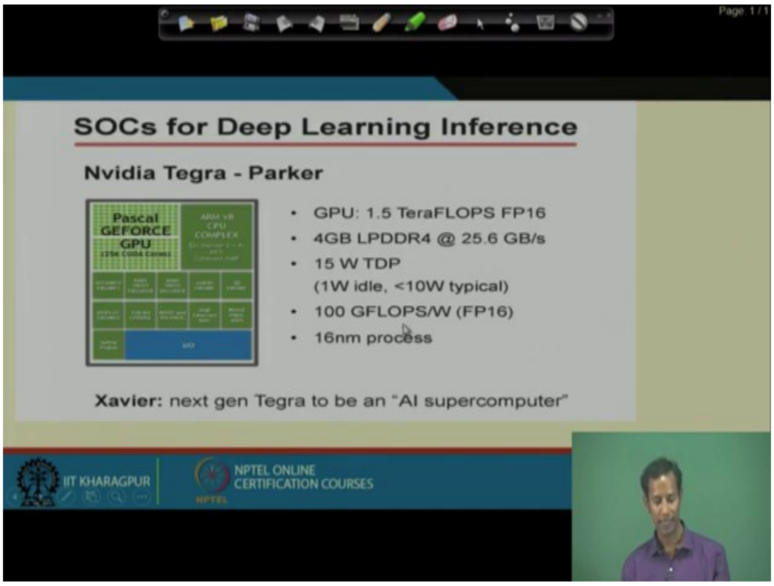
Cloud Systems for Deep Learning

Facebook's Deep Learning Machine

- Open Rack Compliant
- Powered by 8 Tesla M40 GPUs
- 2x Faster Training for Faster Deployment
- 2x Larger Networks for Higher Accuracy

That means all this particular this Deep Learning processes, they are comes up with a huge processing capability ok So, this is one of the Facebook's Deep Learning Machines where it is powered with M 40 GPUs.

(Refer Slide Time: 03:22)



SOCs for Deep Learning Inference

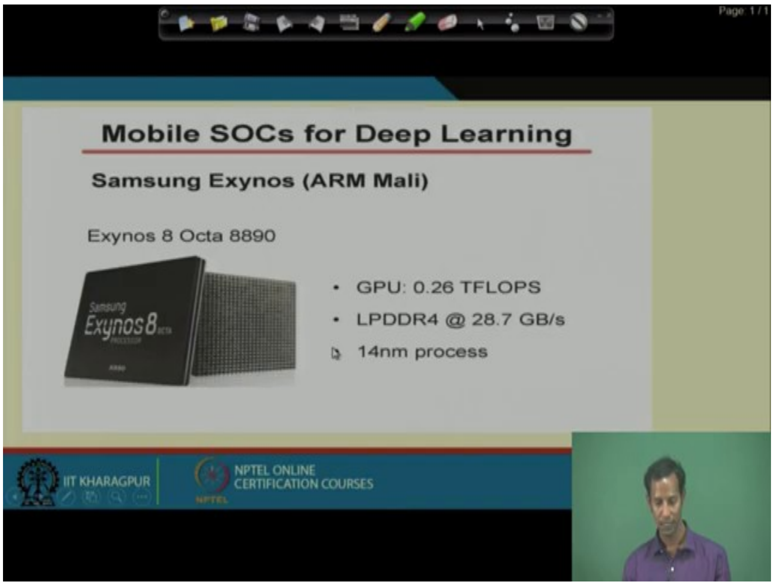
Nvidia Tegra - Parker

- GPU: 1.5 TeraFLOPS FP16
- 4GB LPDDR4 @ 25.6 GB/s
- 15 W TDP
  (1W idle, <10W typical)
- 100 GFLOPS/W (FP16)
- 16nm process

Xavier: next gen Tegra to be an "AI supercomputer"

And then we are having this Nvidia Tegra and in Nvidia Tegra this GPU, consider this 1.5 teraflops ok, and the process it is used which is the 16 nanometer technology.

(Refer Slide Time: 03:39)



Mobile SOCs for Deep Learning

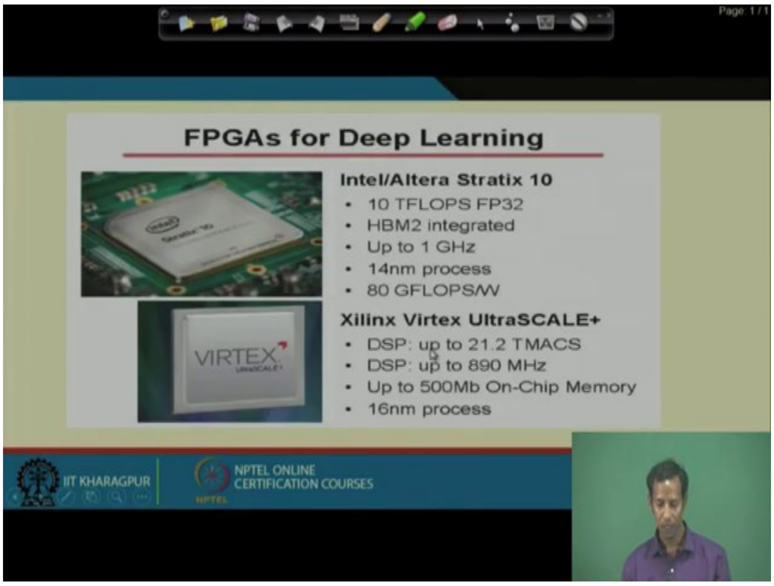
Samsung Exynos (ARM Mali)

Exynos 8 Octa 8890

- GPU: 0.26 TFLOPS
- LPDDR4 @ 28.7 GB/s
- 14nm process

Then we are having this Samsung Exynos which is of this Exynos 8 Octa core processor.

(Refer Slide Time: 03:49)



FPGAs for Deep Learning

Intel/Altera Stratix 10

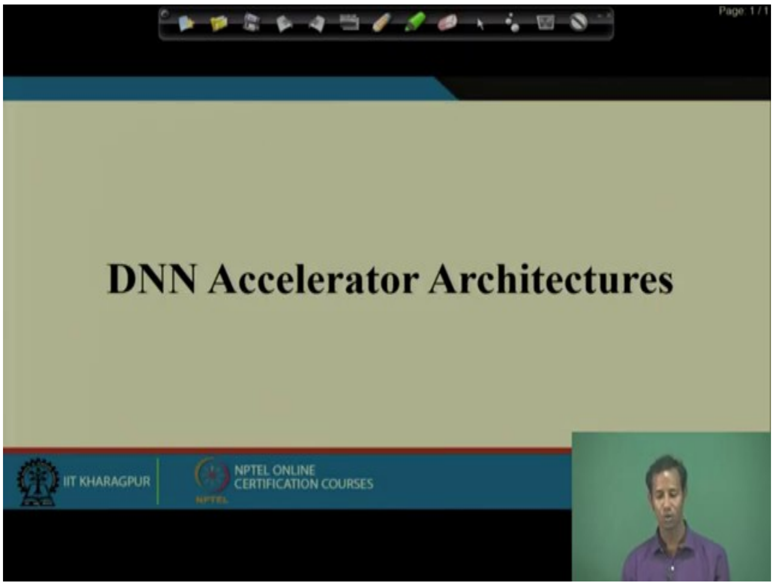
- 10 TFLOPS FP32
- HBM2 integrated
- Up to 1 GHz
- 14nm process
- 80 GFLOPS/W

Xilinx Virtex UltraSCALE+

- DSP: up to 21.2 TMACS
- DSP: up to 890 MHz
- Up to 500Mb On-Chip Memory
- 16nm process

Then we can use this FPGA Xilinx FPGA. These Xilinx, but this Virtex ultra scale it is it you can use this particular FPGA for Deep Learning hardware.

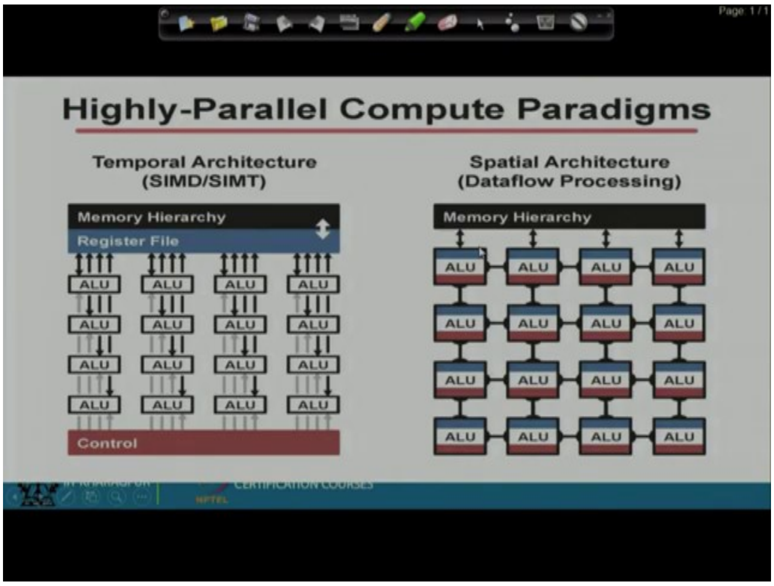
(Refer Slide Time: 04:02)



DNN Accelerator Architectures

So, then actually what is this DNN accelerator architectures look like?

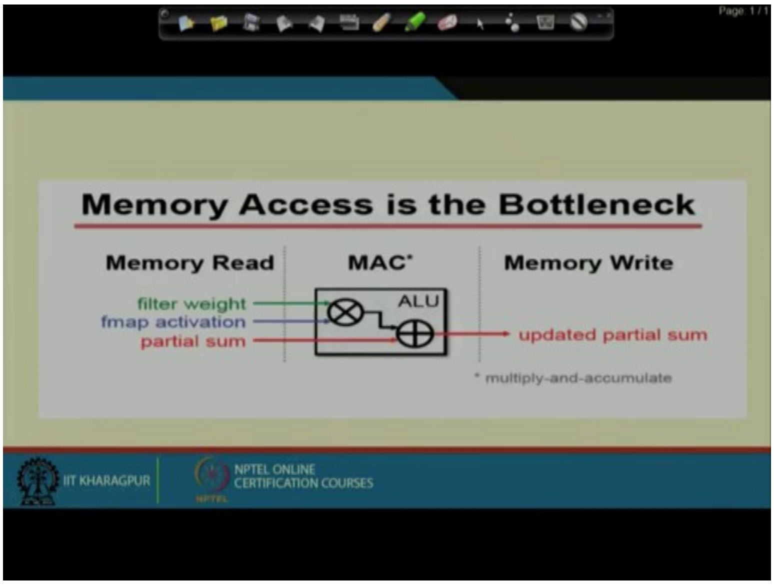
(Refer Slide Time: 04:07)



Highly-Parallel Compute Paradigms

If you see that this is that generic processors architecture where you are having this memory architect memory hierarchy and then you are having this register file where each of these arithmetic logic units, they are basically connected like this. But in case of this machine learning processors, if we have to develop, so at that time we have this ALU to be directly connected with the memory access.

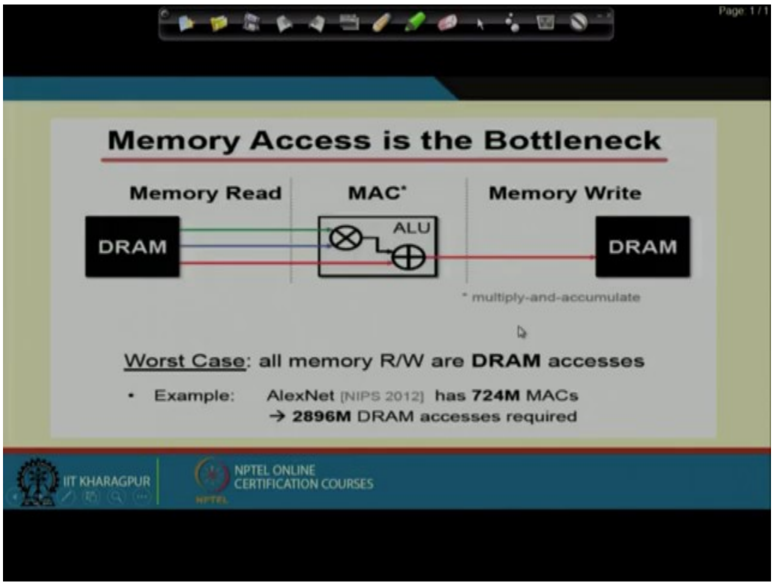
(Refer Slide Time: 04:42)



Memory Access is the Bottleneck

So, why I need actually in this particular for machine learning hardware where memory access is the bottleneck, suppose the ALU which consists of this MAC unit. So, for memory read, I need to read 3 actually component. So, one is this filter weight, another one is this fmap activation and then the partial sum and depending on that the updated partial sum it will be again stored in the memory back.

(Refer Slide Time: 05:17)



Memory Access is the Bottleneck

* multiply-and-accumulate

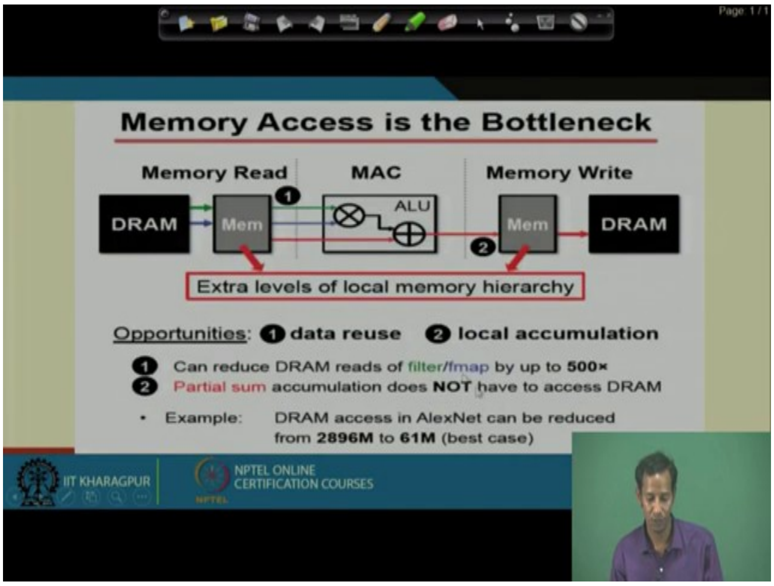
Worst Case: all memory R/W are DRAM accesses

- Example: AlexNet [NIPS 2012] has 724M MACs → 2896M DRAM accesses required

So that means, at that time if I have to train using this AlexNet, which has actually 724 million MAC operations. So, at that time the memory access requirement will be 2896 million number of DRAM access. So, as we know that as the power consumption whenever we access the memory, whether those is for read and write the accessing of the memory consumes more of the power.

So, that is why if actually if we have to develop one; that means, efficient circuit or if you have to close the gap of this the energy requirement in hardware, so at that time we have to reduce the corresponding memory access.

(Refer Slide Time: 06:06)



Memory Access is the Bottleneck

Extra levels of local memory hierarchy

Opportunities: ❶ **data reuse** ❷ **local accumulation**

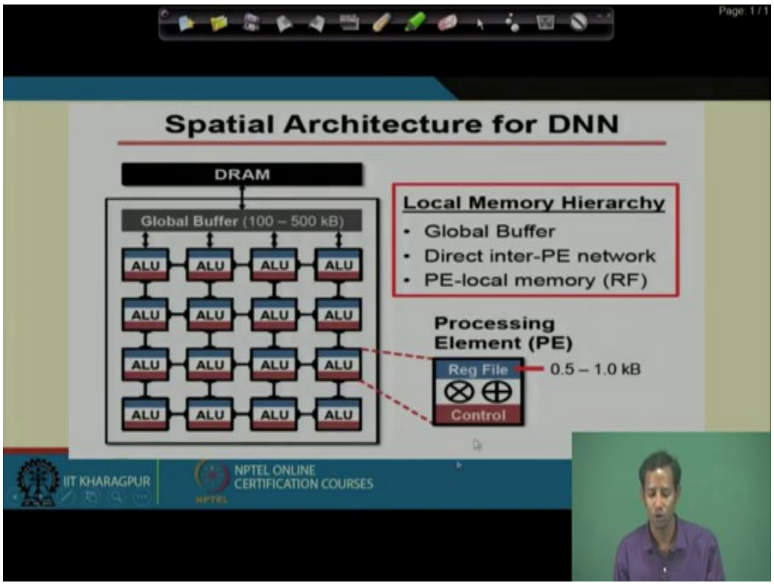
❶ Can reduce DRAM reads of filter/fmap by up to **500×**
❷ Partial sum accumulation does **NOT** have to access DRAM

- Example: DRAM access in AlexNet can be reduced from **2896M** to **61M** (best case)

So, how you can reduce the memory access? So, instead of directly accessing the RAM, if we put some extra levels of local memory at the both for the input side as well as the output side, at that time it can reduce the data by data reusing, the method of data reusing, ok.

So, this data reusing methods it can reduce the D RAM; that means, the power by a factor of 500 x. So, if you just see that from earlier case of which the requirement memory access requirement of was 2896 million, now use of this extra levels of local memory actually it came downs to 61 million number of memory access for this AlexNet, which has which is having 724 millions of mac operation needed for its training of the data sets.

(Refer Slide Time: 07:14)



Spatial Architecture for DNN

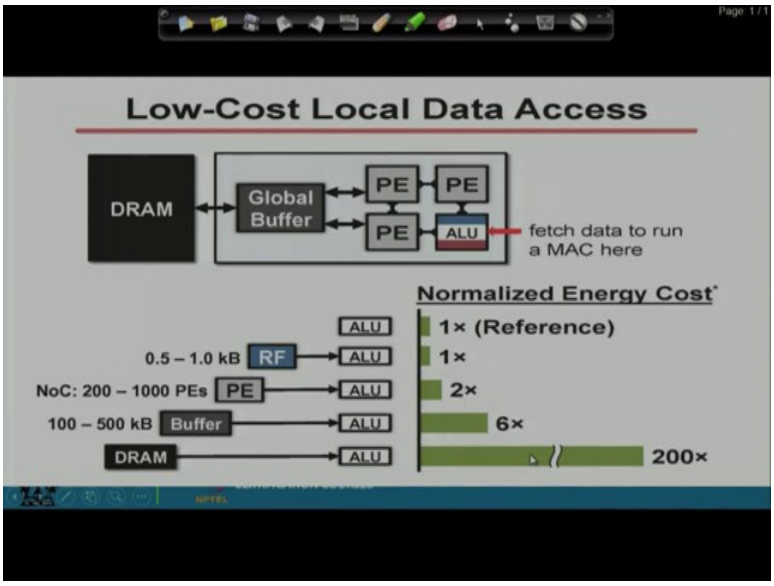
Local Memory Hierarchy

- Global Buffer
- Direct inter-PE network
- PE-local memory (RF)

So, then actually what it actually now what actually this ALU will modified ALU structure will be, each of this ALU will now consider of this local memory hierarchy. So, in the local memory hierarchy, what will be there actually you will be having some DRAMS, then there will be some global buffers and the local memory hierarchy says that you will be needing one global buffers, then direct inter processing elements network that means their processing elements can talk to each other and the processing elements local memory.

So, this the ALU inside it will be something like this there will be the corresponding operations, there will be the control file and along with there will be the register file which will be the size of 0.5 to 1 k, ok. So that means, the inside of each of this ALU, now you need the memory for storing the intermediate values.

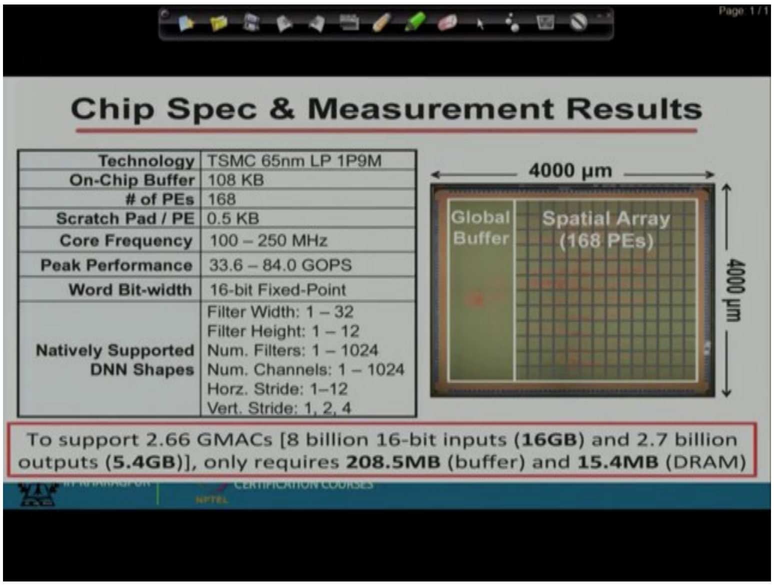
(Refer Slide Time: 08:16)



Low-Cost Local Data Access

So, then if you see that direct access of the ALU to the DRAM it; that means, the energy cost is 200 x times whereas, if we use this global buffer which is the size of 100 to 500 kB, it reduces up to 6 x and when whenever the each of this processing elements; that means, are connected; that means, in this ALU are interconnected then at that time it again reduces that by 2 x terms and if; that means, you are having this local memory to each of this ALU you can just reduce the particular energy cost to a great extent, ok.

(Refer Slide Time: 09:06)



Chip Spec & Measurement Results

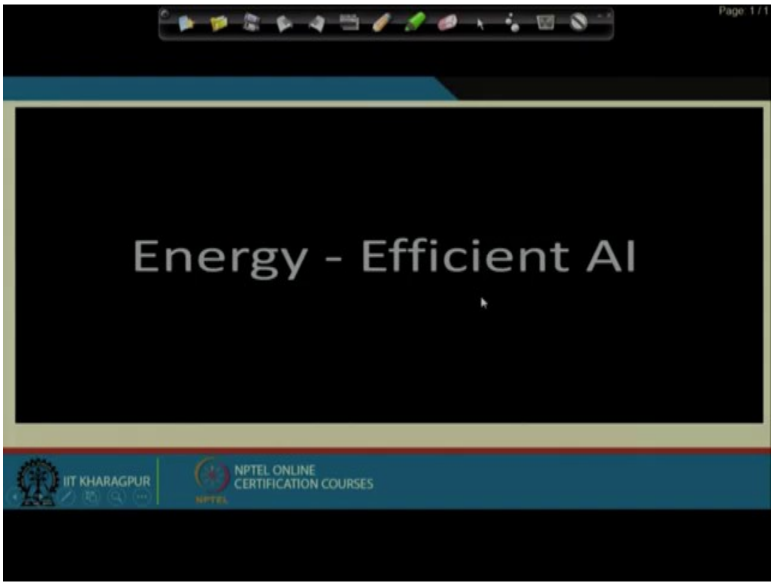
Technology	TSMC 65nm LP 1P9M
On-Chip Buffer	108 KB
# of PEs	168
Scratch Pad / PE	0.5 KB
Core Frequency	100 – 250 MHz
Peak Performance	33.6 – 84.0 GOPS
Word Bit-width	16-bit Fixed-Point
Natively Supported DNN Shapes	Filter Width: 1 – 32 Filter Height: 1 – 12 Num. Filters: 1 – 1024 Num. Channels: 1 – 1024 Horz. Stride: 1–12 Vert. Stride: 1, 2, 4

To support 2.66 GMACs [8 billion 16-bit inputs (**16GB**) and 2.7 billion outputs (**5.4GB**)], only requires **208.5MB** (buffer) and **15.4MB** (DRAM)

So, this is the actually this is 1 this is the chip which is developed based on that whatever actually technology what I said that the ALU which is having this local memory as well as the global memory and they had the each of this process this local that ALU are interconnected to each other and then; that means, after that they have the DRAM access

So, you say you just see that the specification of that particular IC is the it is being developed in 65 nanometer technology, the on chip buffer is 108 KB and the number of processing elements are there in 168 and the processing elements memory is 0.5 KB and then the core frequency is 100 to 250 mega Hertz, the peak performance is 36 point 33.6 to 84 giga operation per second ok So, this is one of the; that means, IC which has been developed by at the research group at MIT.

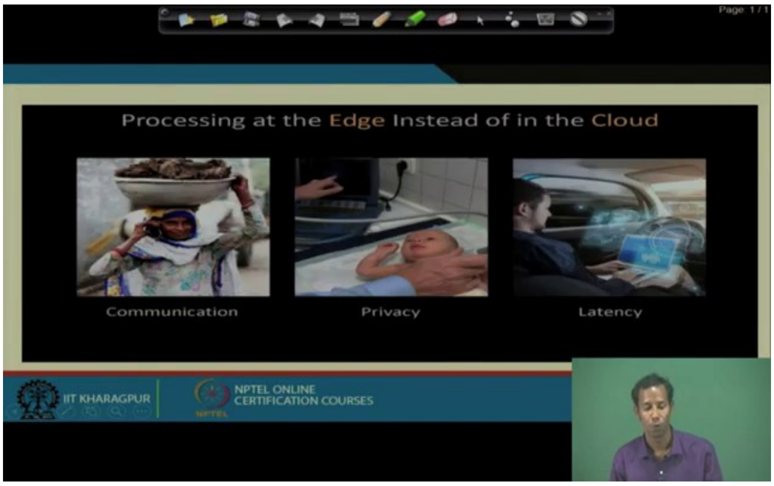
(Refer Slide Time: 10:13)



Energy - Efficient AI

So, now actually we need this energy efficient AI, ok, energy efficient Artificial Intelligence.

(Refer Slide Time: 10:21)


Processing at the Edge Instead of in the Cloud

Why we need energy efficient artificial intelligence is that now, actually we are talking about this cloud computing; what is this cloud computing? Means what actually we are having so much data.

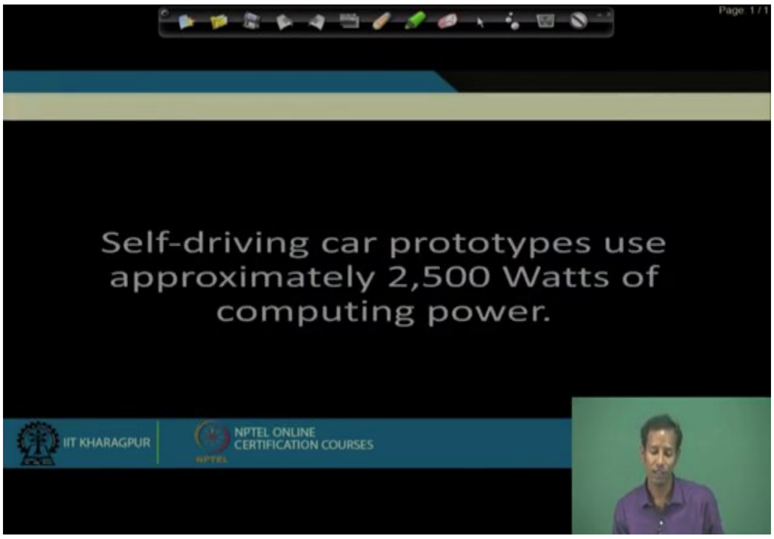
So, now over the cloud you just send to some processing; that means, high processing blocks which are available over the cloud. So, you send the data, process there and then you get back the data. But Edge like for; so instead of, but the recent days actually the requirement of recent days AI that needs the edge computing instead of this cloud computing, why? Because whenever suppose we are actually; that means, the sending the data ok, over the cloud.

So, at that time the security is one of the major issues whenever we process this data in online. The other thing is that suppose I am driving one self driving car. So, at that time suppose one obstacle came. So, at that time to send the data; that means, at that particular instant to send the data over the cloud process and then it will come back and take with the decision it will take too much of time which is the network latency.

So, if the network latency is more at that time, the accident before getting the; that means, the decision or the instruction the accident will occur. So, at that time the oh; that means, the cloud computing is not that much beneficial for that kind of application. So instead, we need one edge computing. Edge computing means the computation will be done at the corresponding terminal or corresponding device or the processor itself. It will not send the data over the network because of this security issue, because of this transmission issues.

So, the in the device itself it will calculate and take the decision immediately so that I can get or I can actually without any accident or without any mishappen mishappening I can just rely on the circuit; that means, this increases the reliability of the artificial intelligence system.

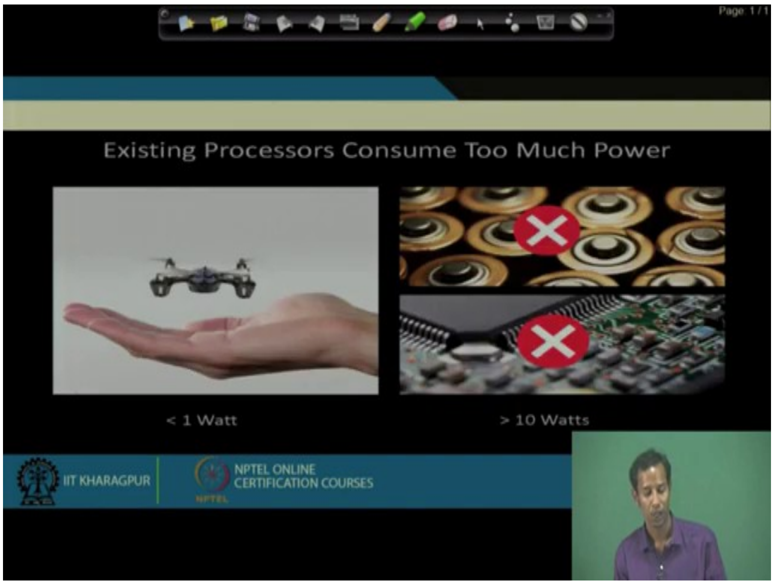
(Refer Slide Time: 12:38)



Self-driving car prototypes use approximately 2,500 Watts of computing power.

So, this one self driving car prototype uses approximately 2500 watts of computing power.

(Refer Slide Time: 12:48)

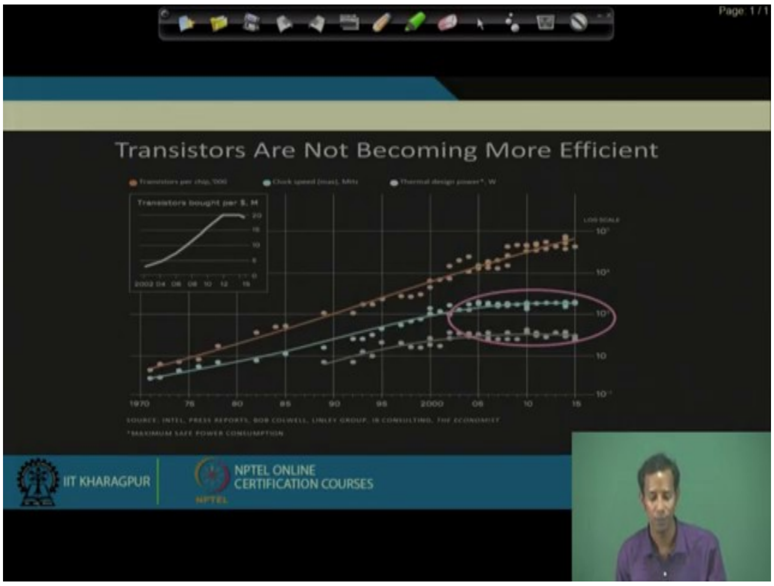


Existing Processors Consume Too Much Power

So, now actually whenever and I said that we are in the age of miniature system design. So, if we see that this is a small drone ok, which has the power budget less than 1 Watt; and nowadays all the systems are miniaturized and they are battery driven. So, the battery driven means their power budget is very much limited.

Now, if we can; that means, within that particular systems, now whenever there is a power budget unless and until I am designing one; that means energy efficient circuit, I cannot provide the artificial intelligence at the edge ok. So, that is why the major challenge in artificial intelligence hardware is that, how actually energy efficient hardware you can make which will support or which will provide this artificial intelligence to the users?

(Refer Slide Time: 13:42)



Transistors Are Not Becoming More Efficient

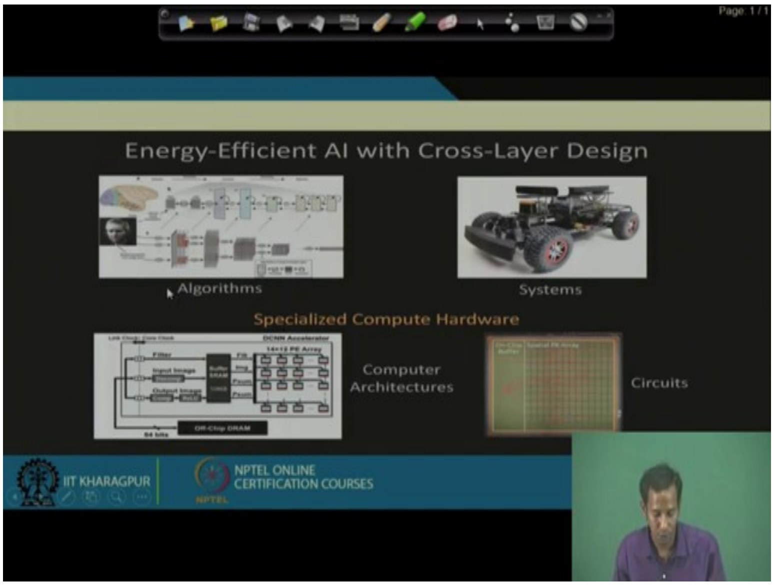
SOURCE: INTEL, PRESS REPORTS, BOB COLWELL, LINLEY GROUP, IB CONSULTING, THE ECONOMIST

*MAXIMUM SAFE POWER CONSUMPTION

So, and then actually if you see the trends in the transistors ok; that means, how these technologies are basically changing?

So, if you see the corresponding transistors are not becoming that much; that means, efficient day by day. So, this is the report which is that got from Intel ok; so, that means, the performance of the transistors are now it is not that day by day it is increasing. It is some kind of it is fixing to somewhere. So, that is why this artificial intelligence with only the help of transistors will not be sufficed to do.

(Refer Slide Time: 14:29)

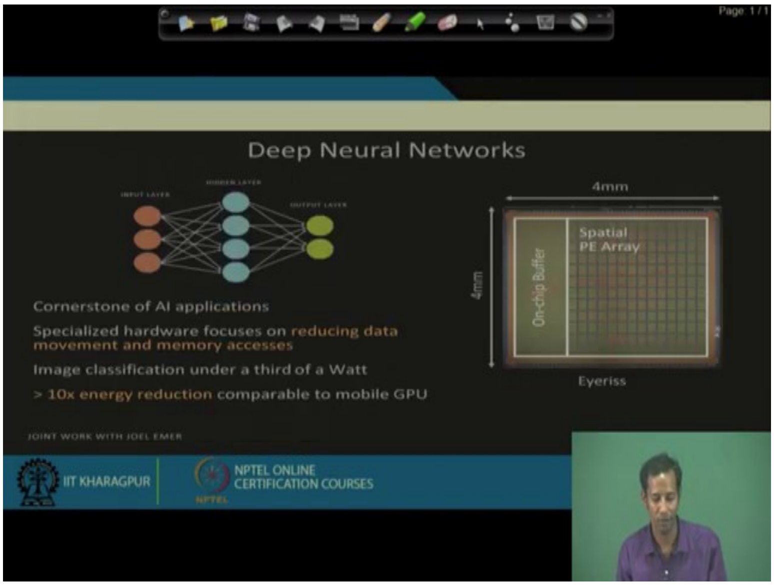

Energy-Efficient AI with Cross-Layer Design

Specialized Compute Hardware

So, at that time what we have to do? We have to create the new algorithm and the new architecture, everything will be the concept will be totally new so that it mainly focus on the energy efficiency or the power efficiency in the hardware. Though the hardware will provide the adaptability or the reconfigurability options, ok. So, we need; that means, innovation at the algorithms level we need the innovation at the system level.

We need the innovation at the circuit level, we need the innovation at the computer architecture level. So that means, all these algorithm, systems, architecture, circuits, everywhere I need innovation. That means, the concept; that means, the generalized actually concept that has to be changed whenever we are focusing to design one artificial intelligence hardware.

(Refer Slide Time: 15:28)



Deep Neural Networks

Cornerstone of AI applications

Specialized hardware focuses on reducing data movement and memory accesses

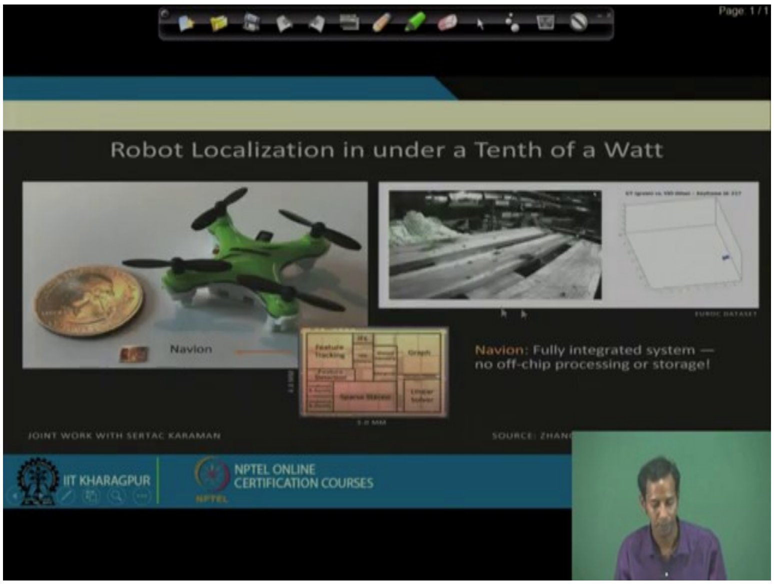
Image classification under a third of a Watt

> 10x energy reduction comparable to mobile GPU

JOINT WORK WITH JOEL EMER

So, this is the deep neural network I have already developed. So this is one of the artificial intelligence processor which is named as Iris, this has been actually developed at MIT. And this work has been published in solid state circuit conference in 2016.

(Refer Slide Time: 15:55)



Robot Localization in under a Tenth of a Watt

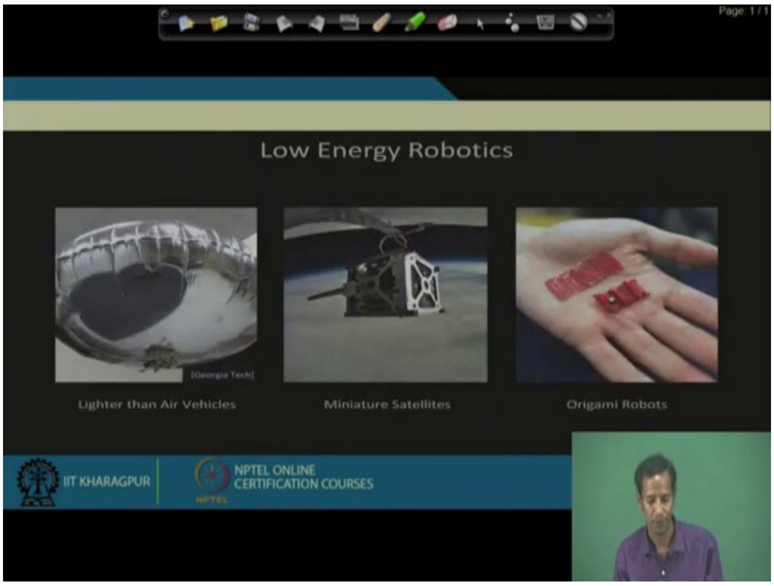
Navion: Fully integrated system — no off-chip processing or storage!

JOINT WORK WITH SERTAC KARAMAN

SOURCE: ZHANG

So, actually we can use this artificial intelligence in robot in; that means, in that means, everywhere. So, if you have heard about the Sophia which is the first humanoid robot. So, it can actually it is having it is the; that means, living example of the artificial intelligence which is basically put inside of the machines.

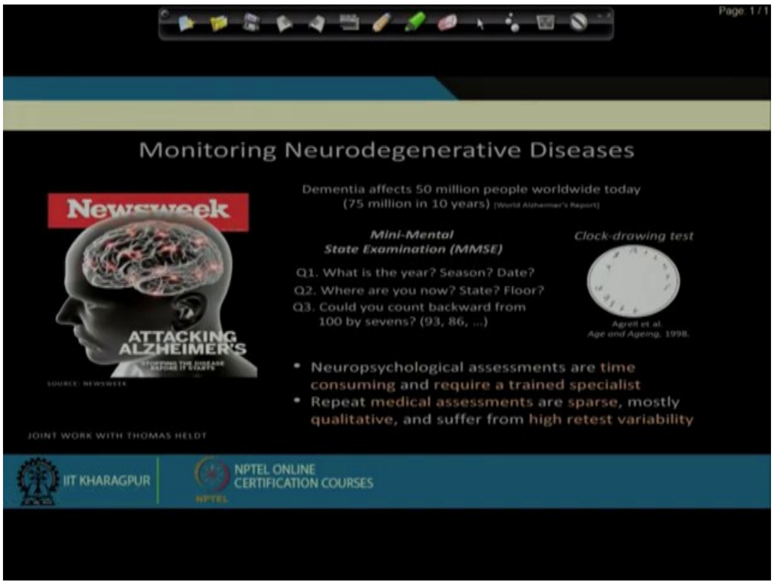
(Refer Slide Time: 16:27)



Low Energy Robotics

So, then actually we needs the low energy robots, so that means, for in this lighter than air vehicles, we need the artificial intelligence support in satellites, we have this origami robots all these are basically low energy robotics application.

(Refer Slide Time: 16:47)



Monitoring Neurodegenerative Diseases

Dementia affects 50 million people worldwide today (75 million in 10 years) (World Alzheimer's Report)

Mini-Mental State Examination (MMSE)

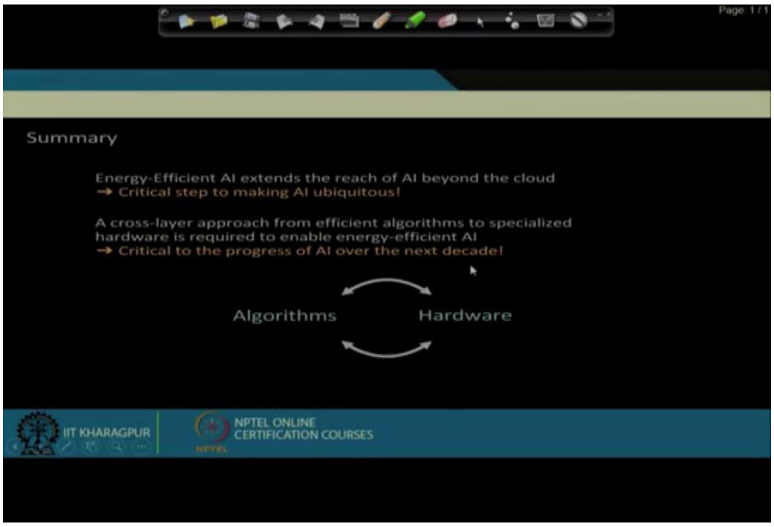
Q1. What is the year? Season? Date?
Q2. Where are you now? State? Floor?
Q3. Could you count backward from 100 by sevens? (93, 86, ...)

- Neuropsychological assessments are time consuming and require a trained specialist
- Repeat medical assessments are sparse, mostly qualitative, and suffer from high retest variability

So, using this AI also I said that this AI can be applicable to the medical application to detect the cancer; that means, to detect the life risking diseases like cancer then the prenatal malfunction also you can you can predict that ok, this if these are the symptoms at that time you are having this.

So, we can use this artificial intelligence we can detect earlier and then we can make you take the correspond; that means, we can take the decision what to do with that patient or we can; that means, we can from the beginning we will be able to know that ok, or we can detect that if there is this problem then this is the outcome of that particular problem.

(Refer Slide Time: 17:40)



Summary

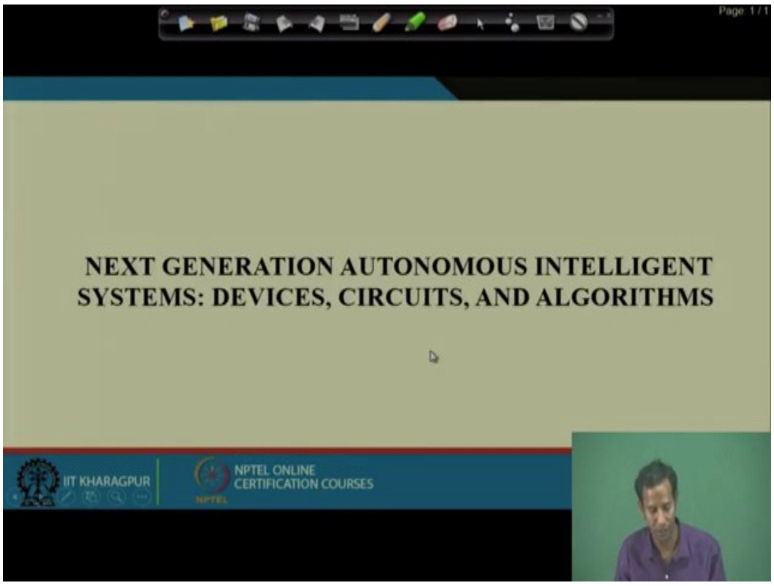
Energy-Efficient AI extends the reach of AI beyond the cloud
→ Critical step to making AI ubiquitous!

A cross-layer approach from efficient algorithms to specialized hardware is required to enable energy-efficient AI
→ Critical to the progress of AI over the next decade!

So, then the summary of this is that the energy efficient AI extend the reach of AI beyond the cloud. So, that is why the all the particular systems requirement is now actually you need the hardware in the era of this edge computing, you need the hardware should be much; that means, it should be energy efficient so that I can get the benefit of the high processing power as well as the low energy level.

So; that means, here if you see we need the innovation or we need there is the challenge and the scope in algorithm development for AI as well as the hardware development for the AI.

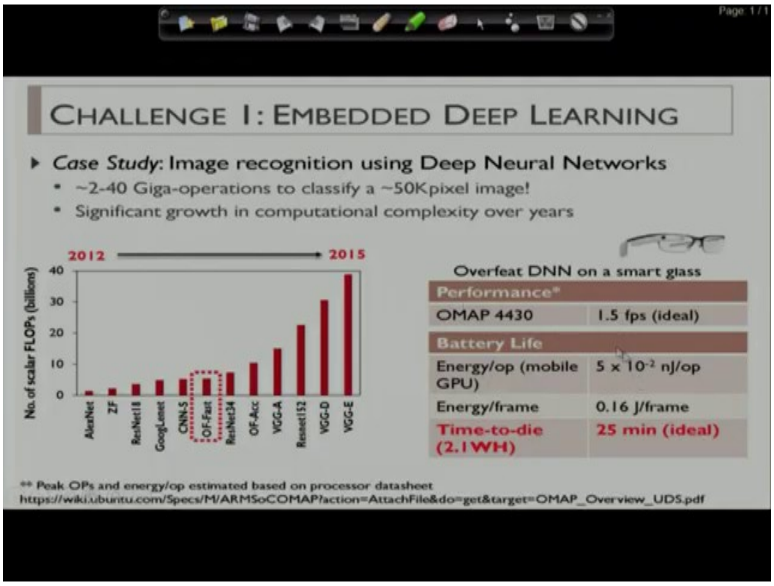
(Refer Slide Time: 18:28)



NEXT GENERATION AUTONOMOUS INTELLIGENT SYSTEMS: DEVICES, CIRCUITS, AND ALGORITHMS

So, the next generation autonomous intelligent system, where we need to focus on the devices the circuits and the algorithms all together.

(Refer Slide Time: 18:38)



Challenge 1: Embedded Deep Learning

- *Case Study*: Image recognition using Deep Neural Networks
  - ~2-40 Giga-operations to classify a ~50Kpixel image!
  - Significant growth in computational complexity over years

Overfeat DNN on a smart glass

Performance*	
OMAP 4430	1.5 fps (ideal)
Battery Life	
Energy/op (mobile GPU)	5×10^{-2} nJ/op
Energy/frame	0.16 J/frame
Time-to-die (2.1WH)	25 min (ideal)

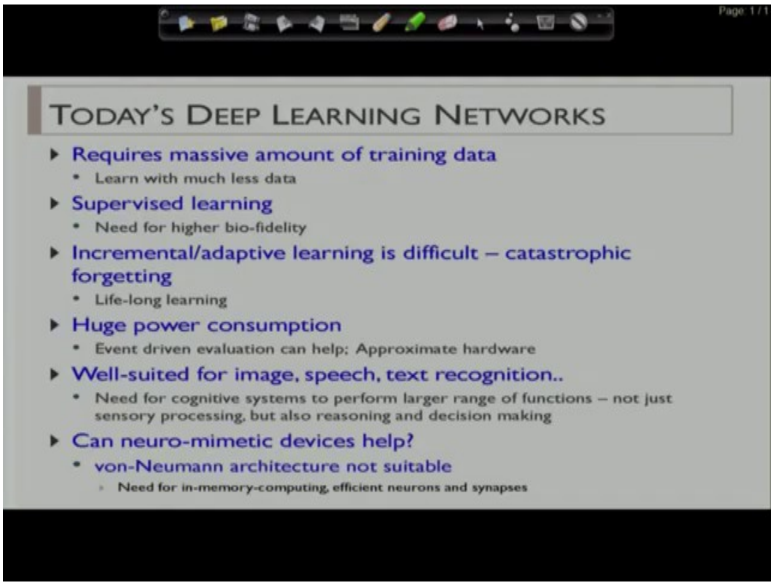
** Peak OPs and energy/op estimated based on processor datasheet
https://wiki.ubuntu.com/Specs/M/ARMSoCOMAP?action=AttachFile&do=get&target=OMAP_Overview_UDS.pdf

So, what is the challenge is that in Embedded Deep Learning actually, we have to suppose we are having using this Google glass, ok.

So, using Google glass I can actually, if I provide the artificial intelligence, so immediately it will recognize and it can be useful for any of the application. So, if you can recognize the object or if you can recognize the person so at that time, you can treat the person properly or you can do anything it will be just like the human being what through his eyes what he can see and he can do or he can take the decision through this Google eyes if I provide artificial intelligence to this particular Google device so, at that time you can do just like a human being.

So, that the major challenge is that the battery life as the battery life it requires a very; that means, small amount of energy which is 5 into 10 to the power minus 2 nano joule per operation ok, in mobile GPU. So, that is why actually you need an energy efficient hardware which will be developed. So, that the power requirement is reduced and you can do the processing high performance processing as well as the power in the lower mode.

(Refer Slide Time: 20:09)



Page: 1 / 1

TODAY'S DEEP LEARNING NETWORKS

- Requires massive amount of training data
  - Learn with much less data
- Supervised learning
  - Need for higher bio-fidelity
- Incremental/adaptive learning is difficult – catastrophic forgetting
  - Life-long learning
- Huge power consumption
  - Event driven evaluation can help; Approximate hardware
- Well-suited for image, speech, text recognition..
  - Need for cognitive systems to perform larger range of functions – not just sensory processing, but also reasoning and decision making
- Can neuro-mimetic devices help?
  - von-Neumann architecture not suitable
    - Need for in-memory-computing, efficient neurons and synapses

So, this is the today's deep learning network, it requires massive amount of training data. If you learn actually with it needs that you need some algorithm which will be developed where you can train your data with less amount of; that means, you can train your network with less amount of data.

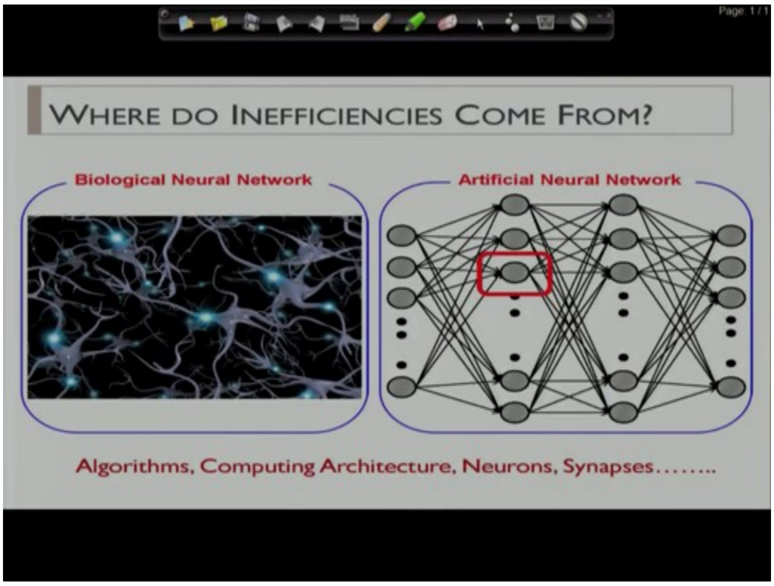
So, less amounted data if you provide to the network for training so, at that time it will be lesser; that means the processing power requirement will be much lesser. So, then you need supervised learning; need for higher bio fidelity, then the incremental or the adaptive learning is difficult; catastrophic forgetting. So, actually as we need actually we are human being as a human being we lifelong we learn.

So, the same thing this lifelong learning also it is needed for this machine learning hardware. And then it needs huge power consumption; so event driven; that means, evolution can help where we can; that means, use this approximate hardware ok, this is one of the technique which is used to reduce the high power consumption along with the techniques what we have already seen.

So; that means, we need that; that means, research or we need the focus to make processing elements, that the processing power at its highest along with the power consumption at its lowest. Then this deep learning network is very much well suited for image, speech and text recognition; and can this neuro-mimetic device help that means, the people are basically trying to develop one new device, ok.

So, the device level, circuit level, system level, algorithmic level, everywhere you are having the challenge so you need to solve those problem.

(Refer Slide Time: 22:28)

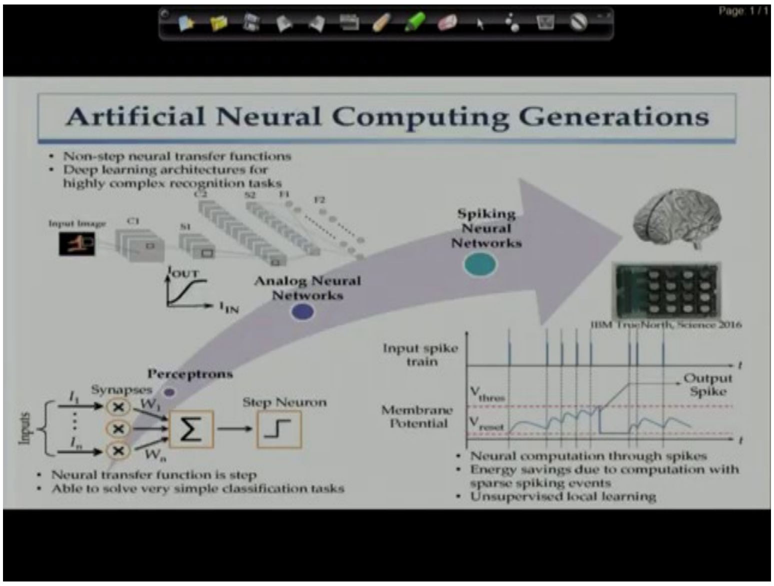


WHERE DO INEFFICIENCIES COME FROM?

Algorithms, Computing Architecture, Neurons, Synapses……..

There are the; that means, there are the problem so, if you can you have to solve the problem. So, that you can get the better results in terms of a good learning in this deep neural network, ok.

(Refer Slide Time: 22:34)



Artificial Neural Computing Generations

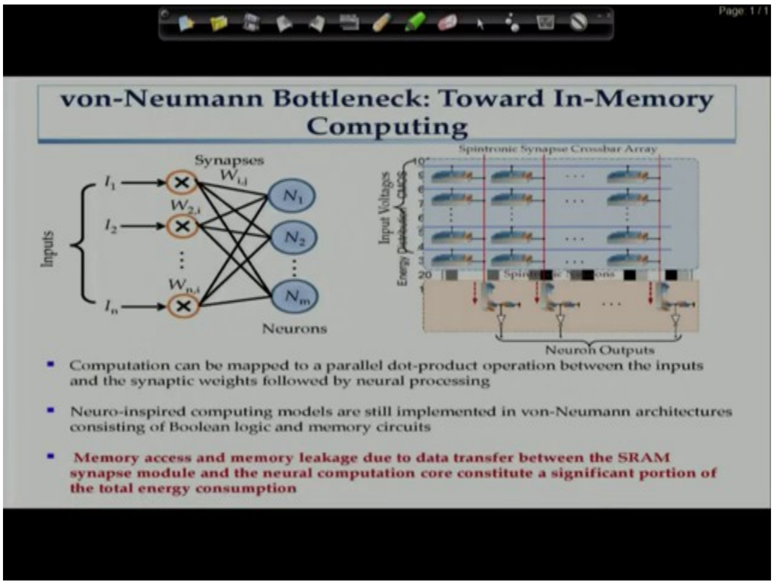
- Non-step neural transfer functions
- Deep learning architectures for highly complex recognition tasks

- Neural transfer function is step
- Able to solve very simple classification tasks

- Neural computation through spikes
- Energy savings due to computation with sparse spiking events
- Unsupervised local learning

So, this is the artificial neural network computing actually, structure; that means, basically what we are trying to do, we are trying to create one artificial brain which will be put inside of the machines ok, and how actually the machines like us we are having the brains which is biological, but the machines will be having this brain not like biological, it will be of mechanical or it will be in terms of circuits, ok.

(Refer Slide Time: 23:10)



von-Neumann Bottleneck: Toward In-Memory Computing

- Computation can be mapped to a parallel dot-product operation between the inputs and the synaptic weights followed by neural processing
- Neuro-inspired computing models are still implemented in von-Neumann architectures consisting of Boolean logic and memory circuits
- **Memory access and memory leakage due to data transfer between the SRAM synapse module and the neural computation core constitute a significant portion of the total energy consumption**

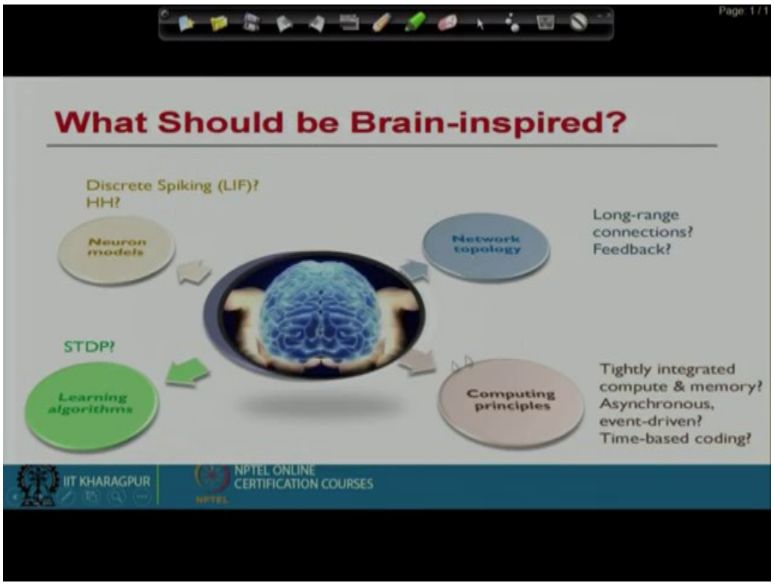
So, this is actually how you can create I already said that you are having this input layer, then you are having this hidden layer and then you are having output layer each of this has been connected with different weight.

And these particular weights will be different or this; that means, the corresponding weight for different classification or different object detection, this will be different. So, the computation can be mapped to a parallel dot product operation, actually this is the earlier von-Neumann architecture which is used in computer architecture. So, the bottleneck is that so which leads to this in-memory computing. Today people are basically to support this AI; they are providing this in-memory computing.

The earlier IRISH also whatever we have discussed, there also the each of this ALU contents, some of these local memory, then they have this global buffers then they have this the access to the DRAM, ok. So, those as basically reduce the write and read of the memory so by that they can save the energy. So, the generic von-Neumann architecture; the processor architecture, has the bottleneck for this memory access which can consumes a lot of power.

So, that is why we need in future we need this in-memory computing which will be much more efficient for artificial intelligence hardware development ok.

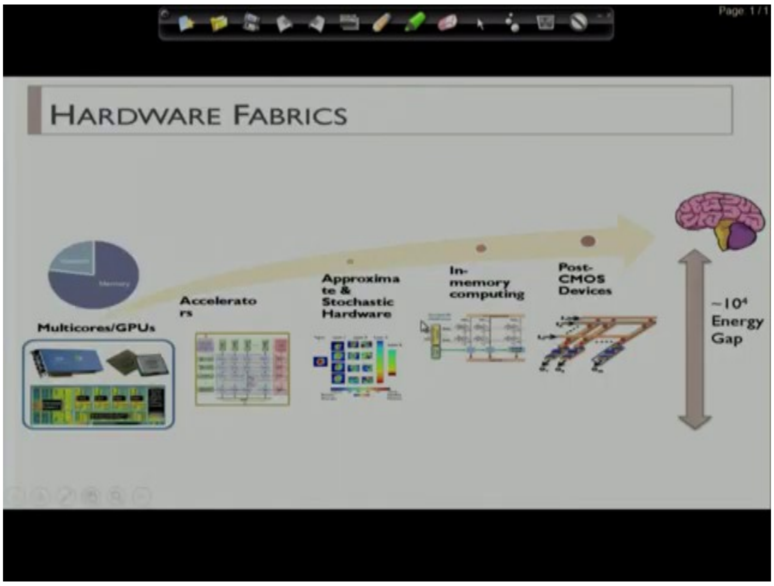
(Refer Slide Time: 24:51)


What Should be Brain-inspired?

So, if we just see this brain. So, we have to develop this neuron models, we have to develop this learning algorithms, we have to develop this network topologies and we have to set the computing principles ok.

So, whenever we will complete all this together, then we can create one artificial brain which will be known as artificial intelligence ok.

(Refer Slide Time: 25:24)



HARDWARE FABRICS

Multicores/GPUs

Accelerators

Approximate & Stochastic Hardware

In-memory computing

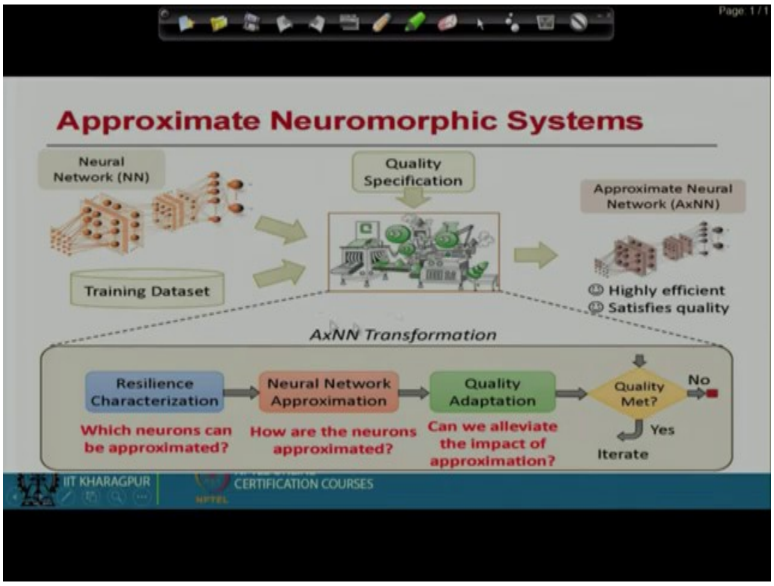
Post-CMOS Devices

~10^4 Energy Gap

So, the hardware fabrics actually you can use this multicore GPUs, you can use these accelerators, you can use this approximate and stochastic hardware’s, you can use this in-memory computing, you can do this post CMOS devices or changes in the CMOS devices to get; that means, the device itself will be having the corresponding memory itself.

So, to all this, actually opens up the research scope; by what? To close this energy gap where is requirement of 10 to the power 4 with respect to our normal brain to the artificial brain. So, that is why actually there is the scope or there is the options, now whatever we have learned in this particular course that techniques or those whatever learning things we have done so those things we can apply to make this artificial intelligence hardware more efficient energy efficient in future, ok

(Refer Slide Time: 26:43)



Page: 1 / 1

Approximate Neuromorphic Systems

Neural Network (NN)

Training Dataset

Quality Specification

Approximate Neural Network (AxNN)

☺ Highly efficient
☺ Satisfies quality

AxNN Transformation

Resilience Characterization → Neural Network Approximation → Quality Adaptation → Quality Met?

Which neurons can be approximated?

How are the neurons approximated?

Can we alleviate the impact of approximation?

No

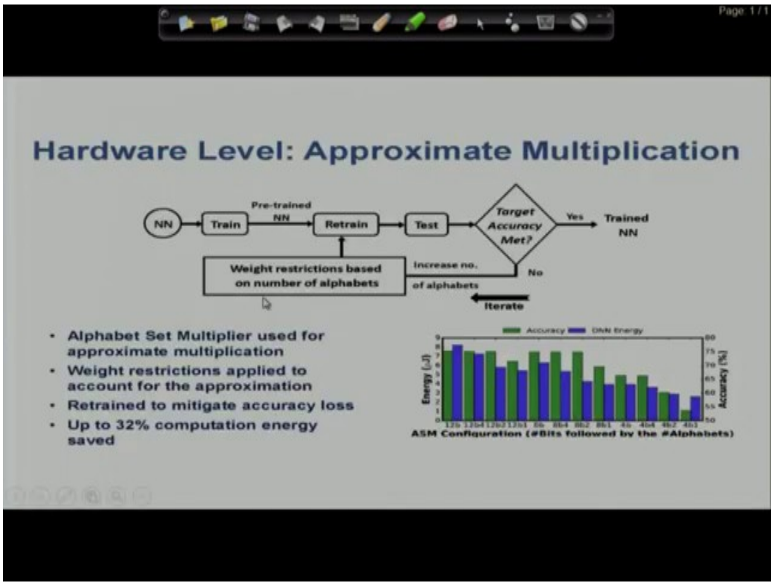
Yes

Iterate

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is Approximate Neuromorphic System, this work is following in Purdue University.

(Refer Slide Time: 26:55)



Page: 1 / 1

Hardware Level: Approximate Multiplication

NN → Train → Pre-trained NN → Retrain → Test → Target Accuracy Met? → Yes → Trained NN

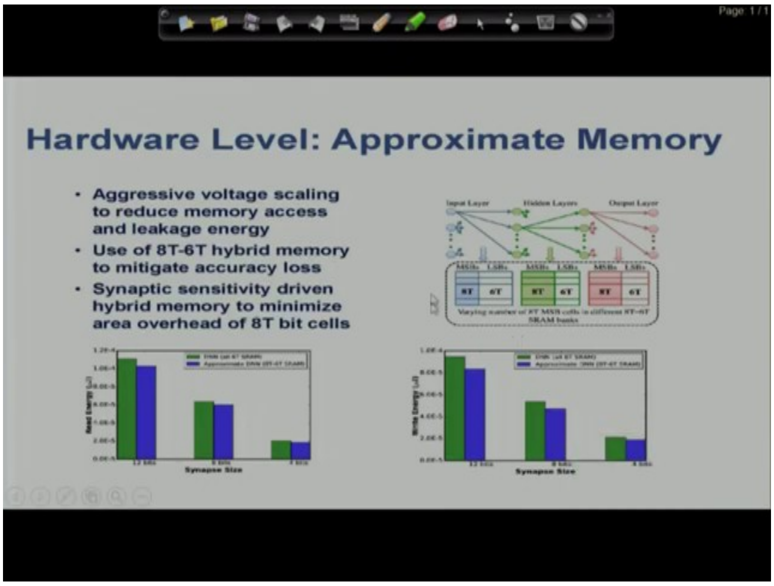
No → Increase no. of alphabets → Weight restrictions based on number of alphabets → Retrain

Iterate

- Alphabet Set Multiplier used for approximate multiplication
- Weight restrictions applied to account for the approximation
- Retrained to mitigate accuracy loss
- Up to 32% computation energy saved

So, and then using this approximate architecture hardware architecture, you can reduce the hardware; that means, you can reduce the hardware as well as you can reduce the power consumption too. Where, you are basically compromising the with the accuracy ok.

(Refer Slide Time: 27:12)

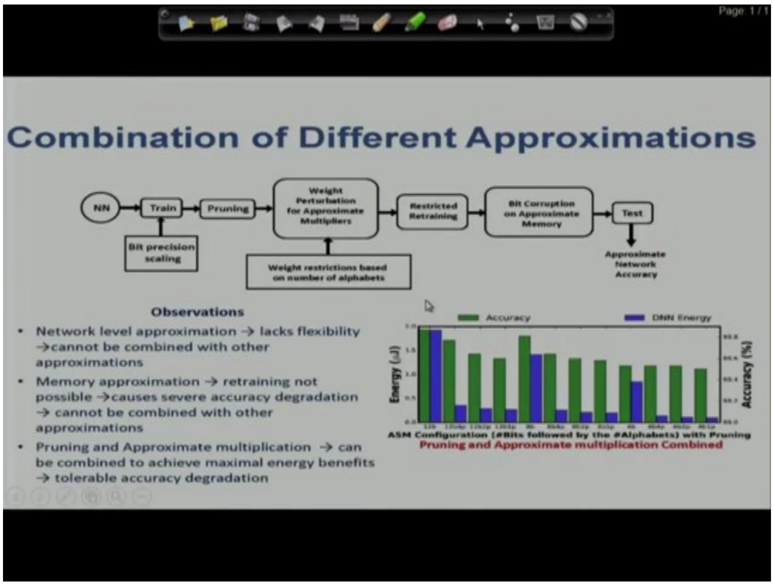


Hardware Level: Approximate Memory

- Aggressive voltage scaling to reduce memory access and leakage energy
- Use of 8T-6T hybrid memory to mitigate accuracy loss
- Synaptic sensitivity driven hybrid memory to minimize area overhead of 8T bit cells

So, then this is approximate memory.

(Refer Slide Time: 27:16)



Combination of Different Approximations

Page: 1 / 1
NN
Train
Pruning
Weight Perturbation for Approximate Multipliers
Restricted Retraining
Bit Corruption on Approximate Memory
Test
Bit precision scaling
Weight restrictions based on number of alphabets
Approximate Network Accuracy
Accuracy
DNN Energy
Energy (uJ)
Accuracy (%)
ASM Configuration (#Bits followed by the #Alphabets) with Pruning

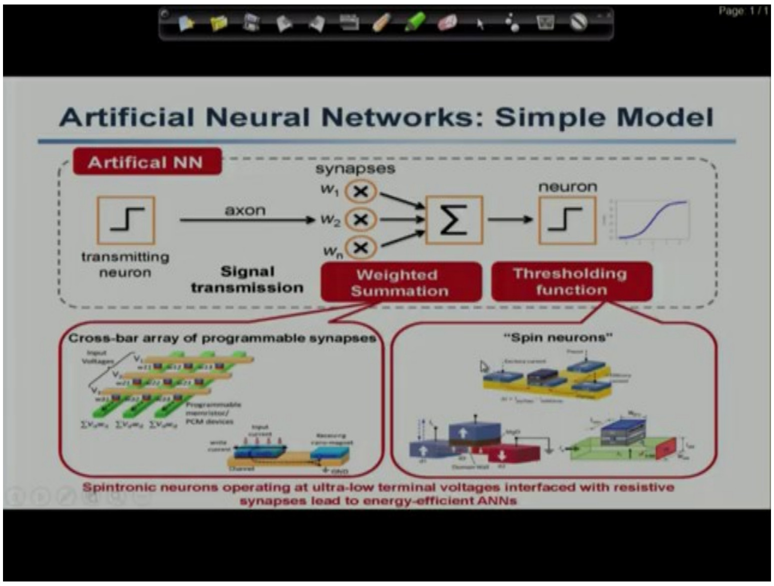
Observations

- Network level approximation → lacks flexibility →cannot be combined with other approximations
- Memory approximation → retraining not possible →causes severe accuracy degradation → cannot be combined with other approximations
- Pruning and Approximate multiplication → can be combined to achieve maximal energy benefits → tolerable accuracy degradation

Pruning and Approximate multiplication Combined

And then the; that means, each of this particular model; that means, the overall deep neural network has been developed using this approximation at different blocks. So, whenever we; that means, use this approximation hardware. So, at that time the accuracy I have to compromise there at the accuracy, but the DNN energy is significantly reduced if we use this approximate hardware for artificial intelligence.

(Refer Slide Time: 27:52)

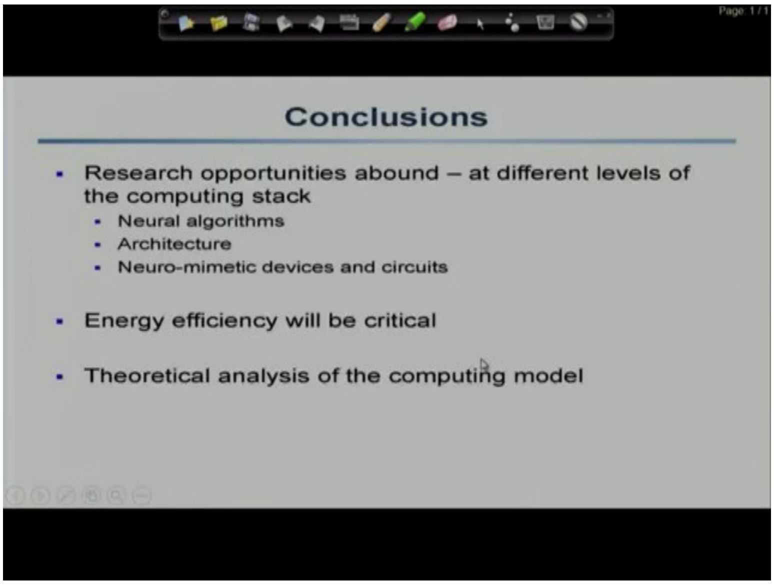


Artificial Neural Networks: Simple Model

Spintronic neurons operating at ultra-low terminal voltages interfaced with resistive synapses lead to energy-efficient ANNs

So, then again the other option is that this in-memory computing people are trying to do this in-memory computing then this spintronic devices have been developing. So, using this spintronic devices, in the device level itself you can; that means, have the memory options. So, it is just like the neurons which is which store the situation, the same thing you can put in the device level by this spintronic devices.

(Refer Slide Time: 28:25)



Conclusions

- Research opportunities abound – at different levels of the computing stack
  - Neural algorithms
  - Architecture
  - Neuro-mimetic devices and circuits
- Energy efficiency will be critical
- Theoretical analysis of the computing model

So, there are this research opportunities around at different levels of computing stack; whether that is in this neural network algorithms or whether there the architectures of these VLSI architectures or this neuro-mimetic devices and circuits and everywhere the primary task is that energy efficiency is the critical part of artificial intelligence hardware design. And then theoretical analysis of the computing model then the algorithm which will take lesser amount of training data to properly train the hardware or train the machines.

So, all this creates the scope to work on this machine learning hardware and again I said the machine learning hardware is not application specific, it is more like adaptive kind of things or adaptive hardware we have to design. So, with this particular lecture or with this vision that if you are more interested in machine learning hardware design. So, at that time I think that this course will help you a lot.

Because actually we have learned the techniques how to achieve or how to get this; that means, low power or how to get this high speed or how to get this low area; at the architectural level. So, if you are interested in this that hardware architecture which will be developed for machine learning, then you can use this particular course, you can gain the experience from this course and then you can apply for developing the upcoming or the emerging artificial intelligence hardware design. So, this is it and.

Thank you.