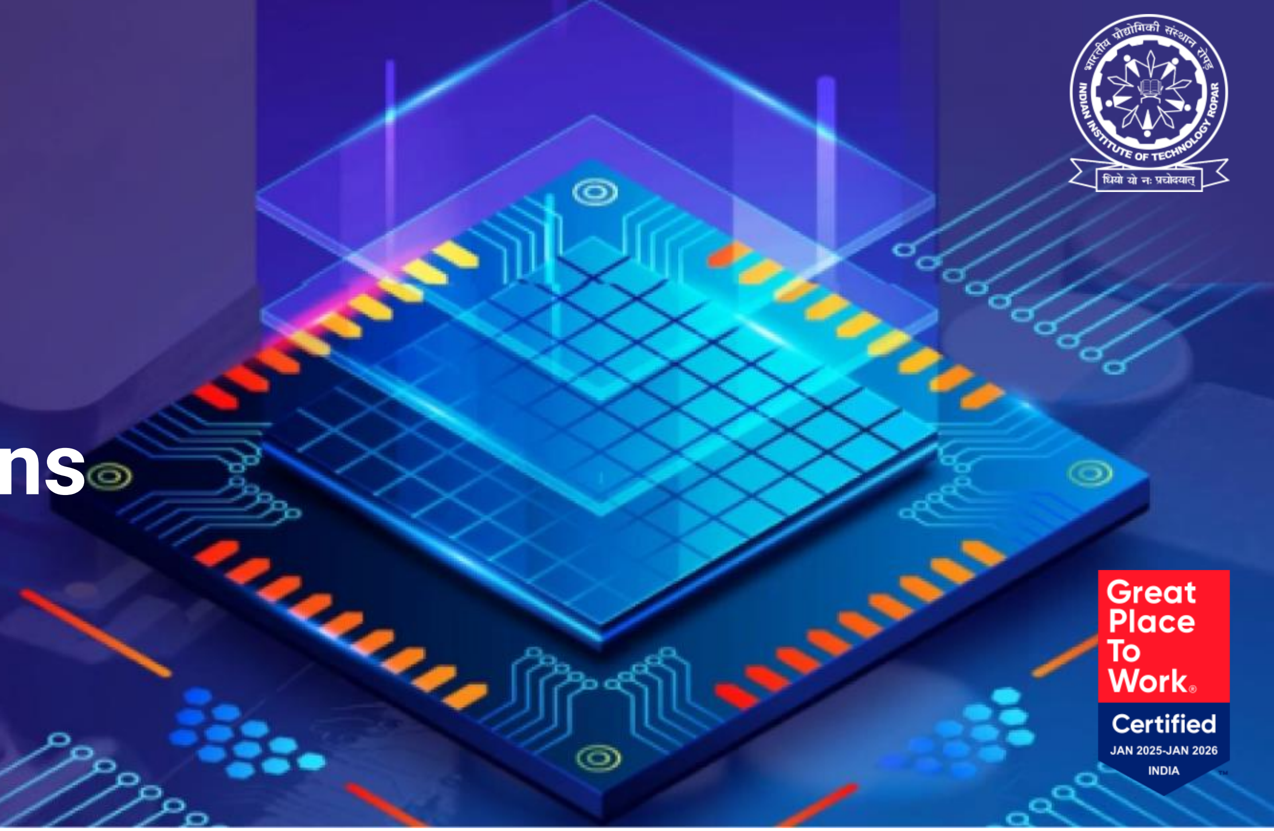


Your trusted Global Partner in Proven Semiconductor Solutions

AI & Semiconductor Design

Dr. Neeraj Goel - Speaker

Associate Professor, CSE,
IIT Ropar

Guniyal Bagga - Host

Associate Brand Marketing,
Orbit & Skyline

Introduction

About Our Customers

Agenda

Overview of Webinar Structure

01 AI/Machine learning premier

02 Use of AI in Semiconductor design

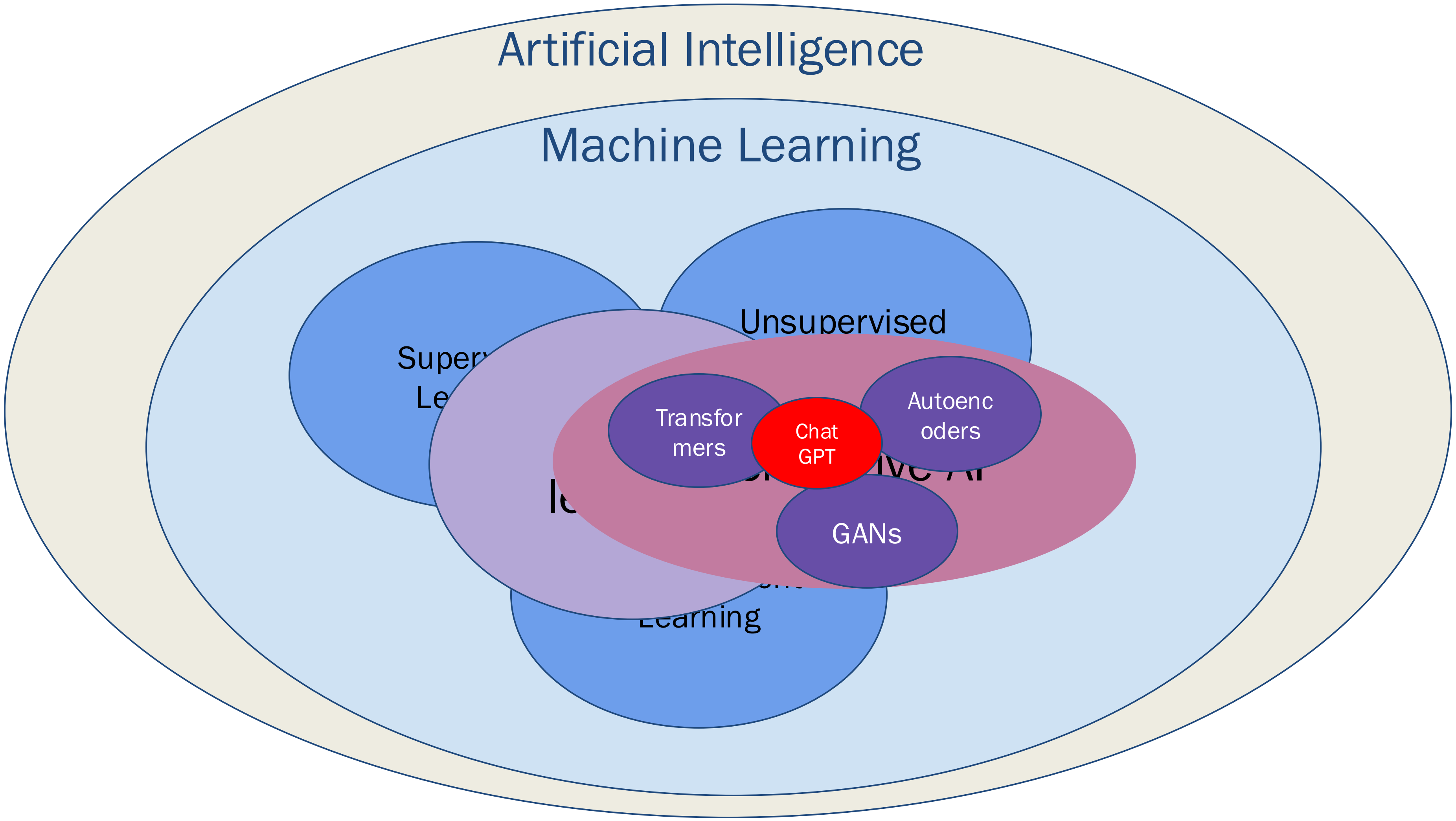
03 Chip design and its impact on AI

04 AI and Embedded/edge computing

05 Q & A

06 Semicon India 2025

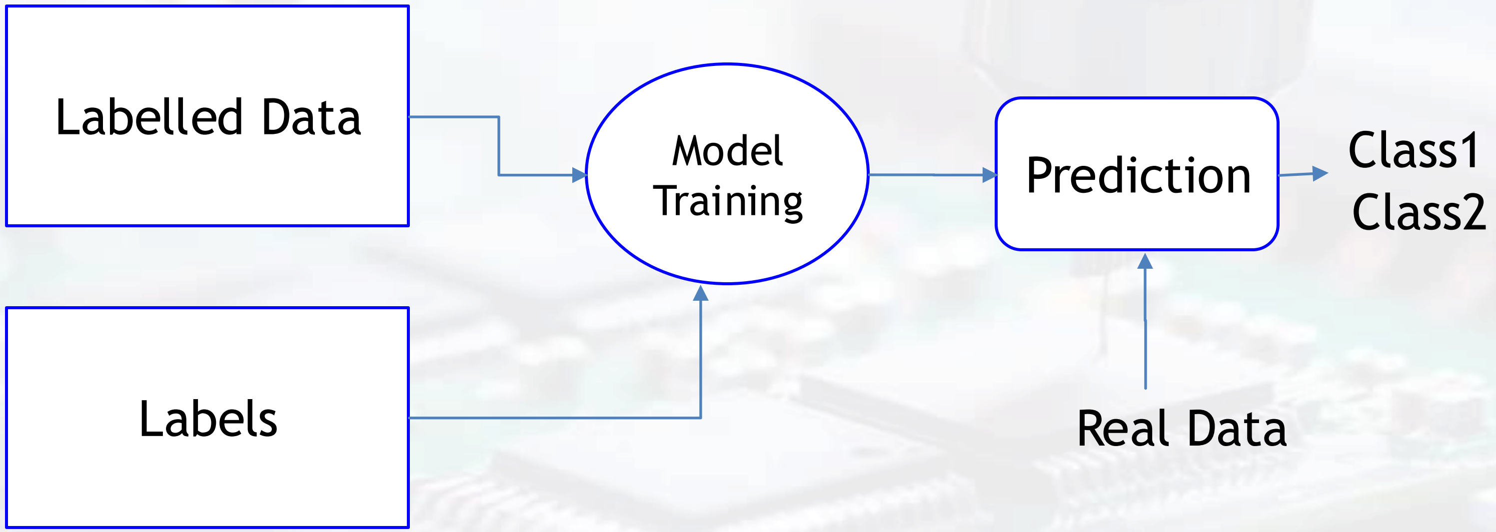
AI Landscape and Terminology



Machine Learning - Basics

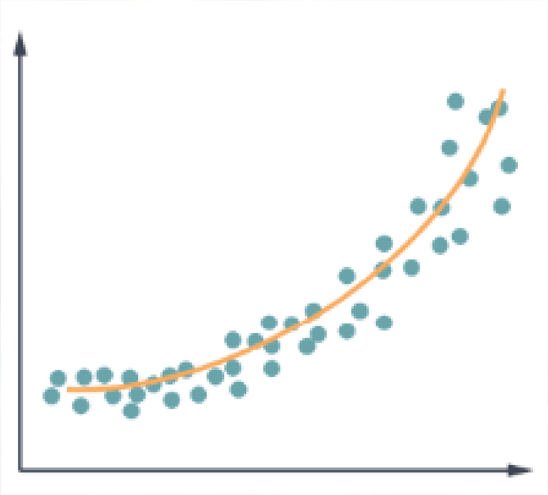
- Supervised
  - Classification

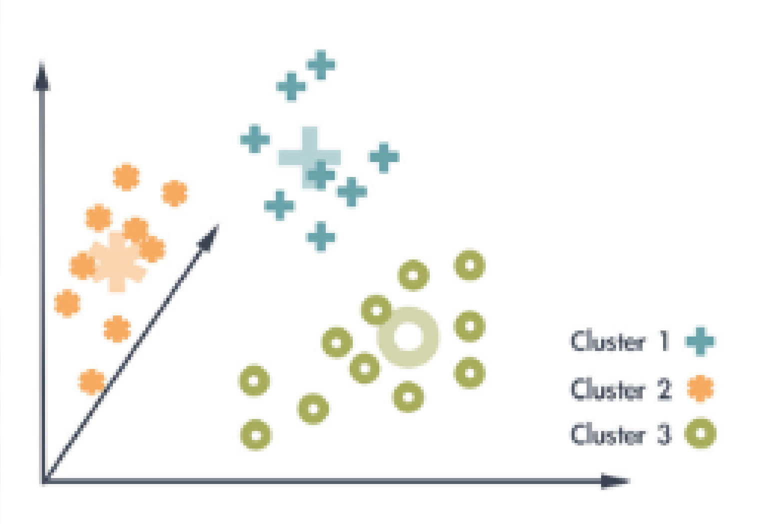
- SVM
- Perceptron
- Random-forest
- Decision Tree
- ANN
- Deep learning
- LSTM
- Transformers



Machine Learning - Basics

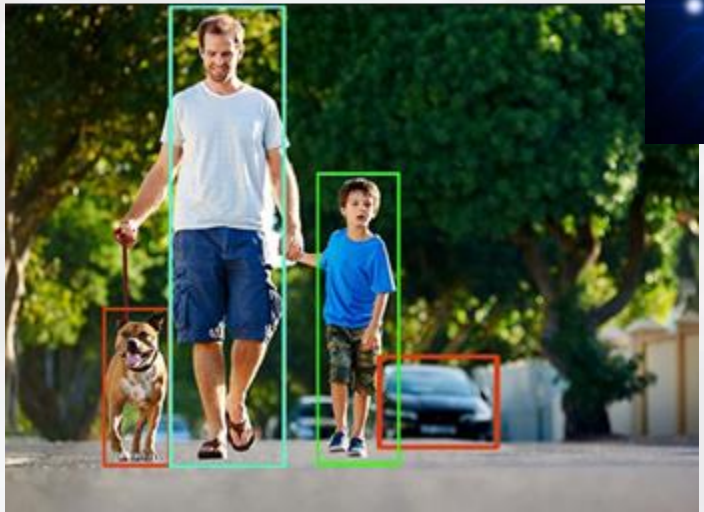
- Supervised
  - Classification
  - Regression
- Unsupervised
  - Clustering

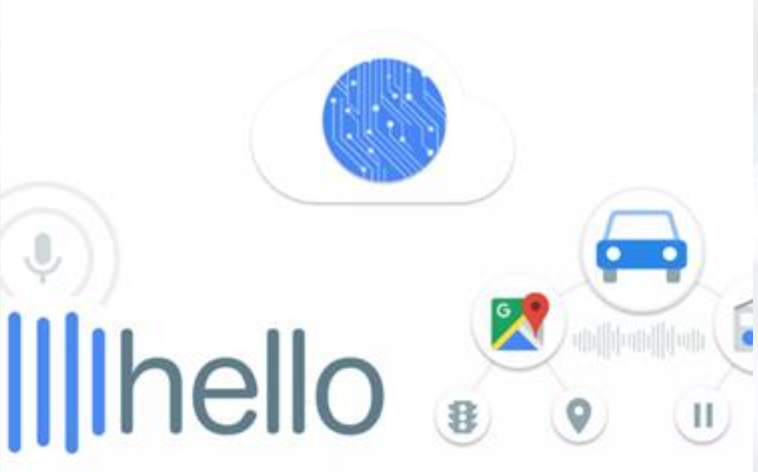




Summary:
- Given training data, ML algorithms can predict class of unknown data
- Can model complex data behaviour

Artificial Intelligence Applications

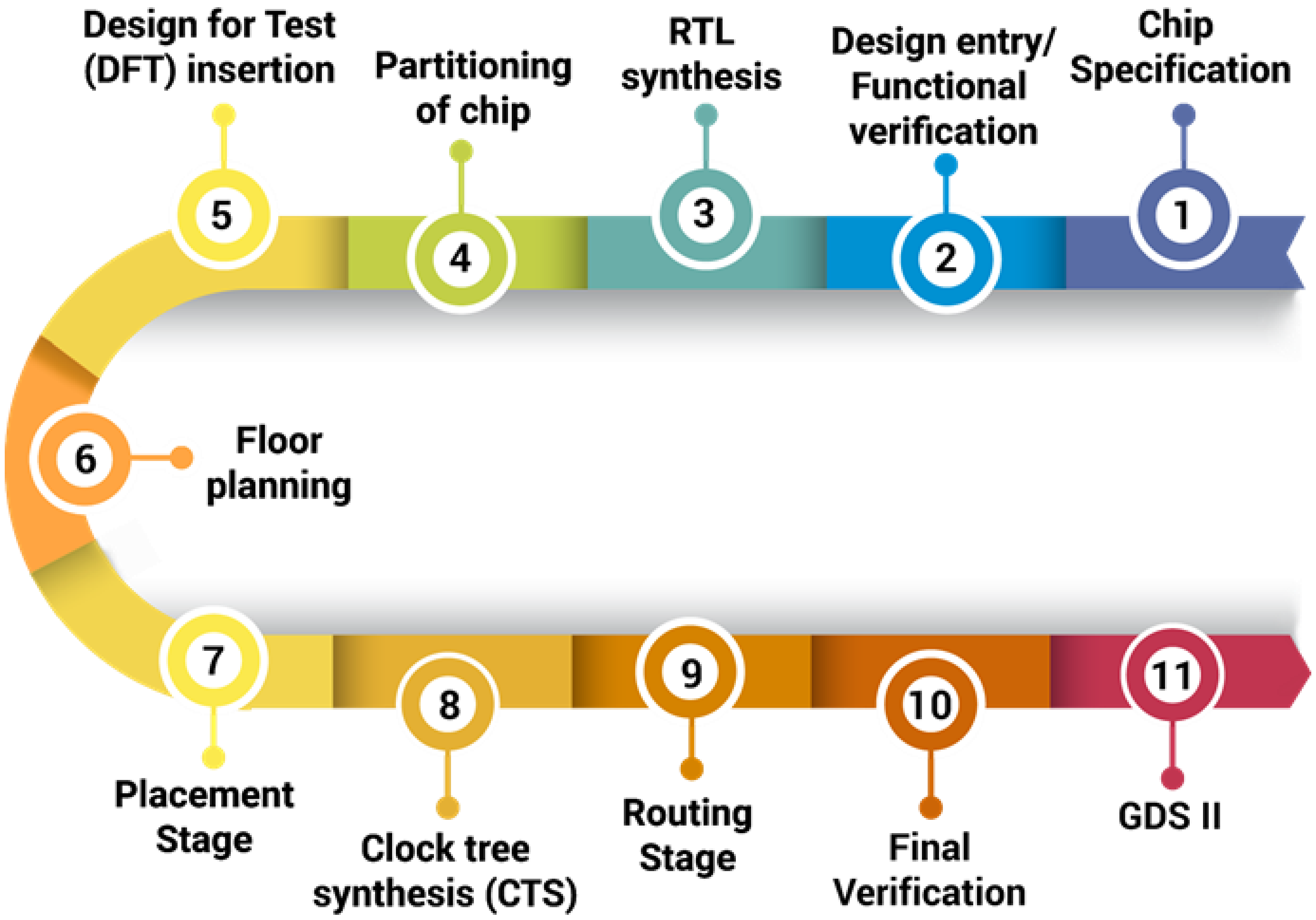




source: google images

Use of AI in Semiconductor design

Semiconductor Design Process



Challenges

- Long design cycle
  - Engineering time
  - EDA is slow
- # transistors
- Iterative
- EDA algorithms NP complete

Image source: https://www.einfochips.com/blog/

How AI or ML is Helping in Semiconductor Design

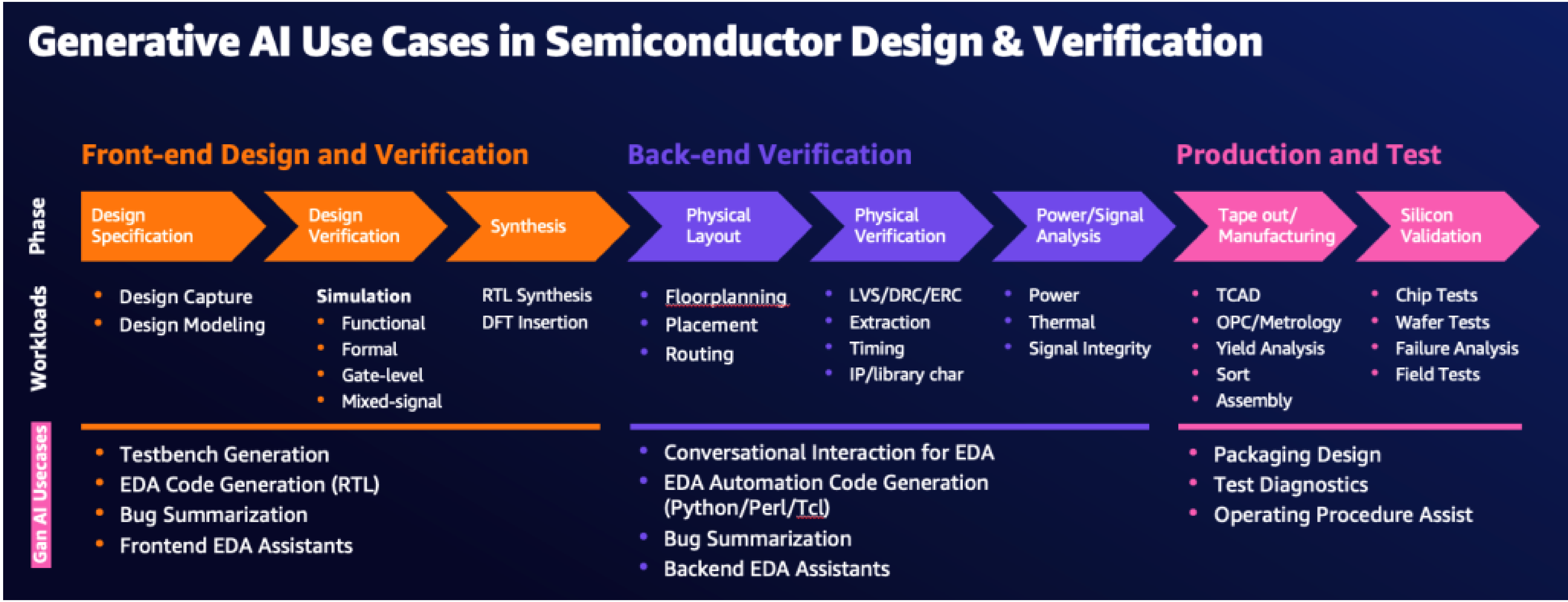
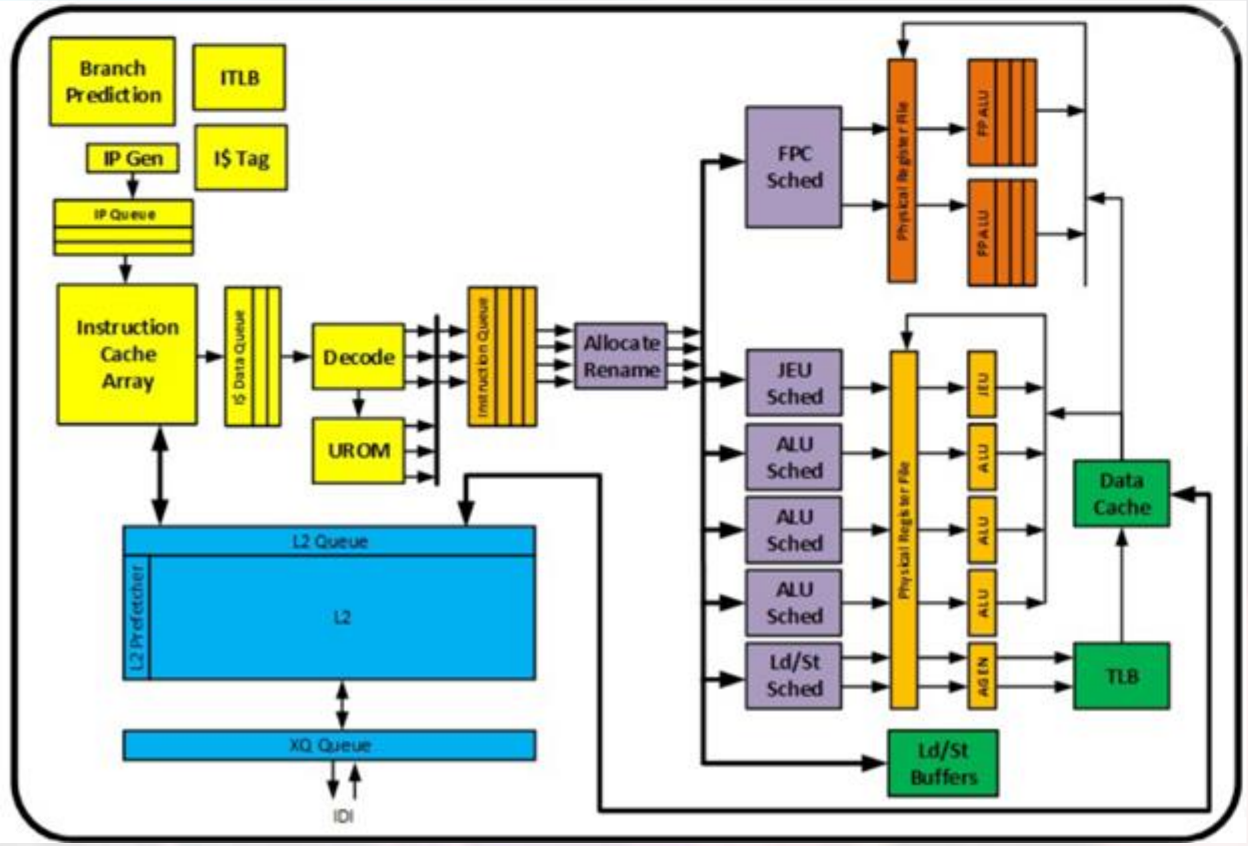


Image source: AWS Blogs

ML in Computer System Design

- **Design of Processor core**
  - Branch prediction
  - Custom instructions
  - Instruction scheduling
- **Design of memory sub-system**
  - Prefetch
  - Cache: replacement policy, cache partitioning
  - Cache: Set utilization
  - Non-volatile memory
- **Multi-processors**
- **Different workloads**

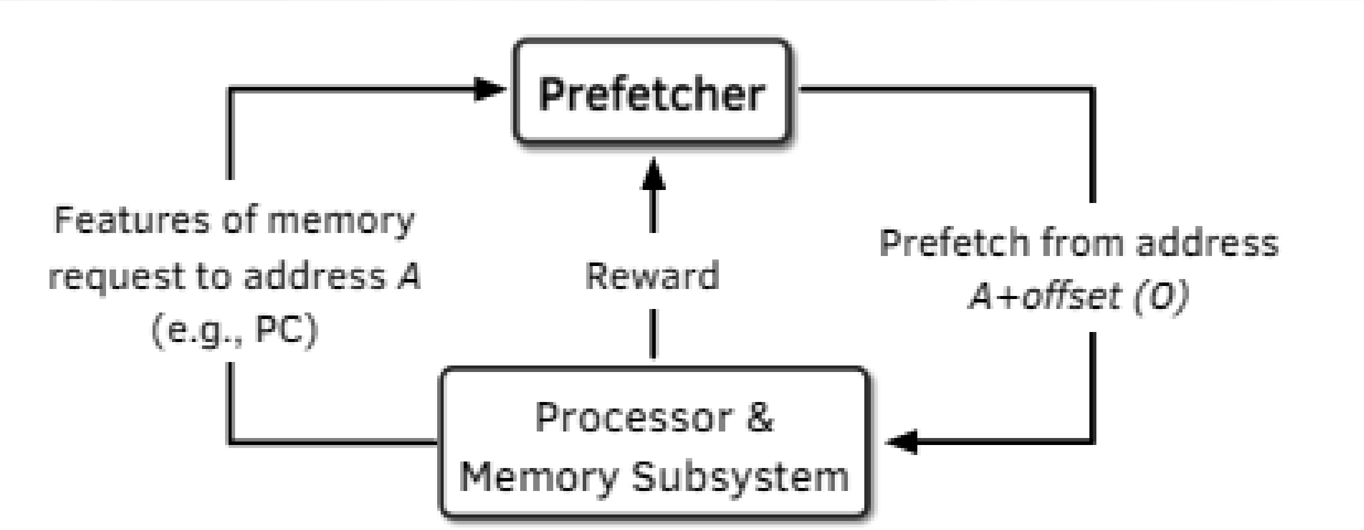


Prefetching

- Question: Which address to prefetch, when to prefetch (spatio-temporal locality)
- Conventional methods: Calculating strides distances
- ML methods:
    - As classification problem or regression problem
    - LSTM: Long warmup and prediction latency

Reinforcement Learning in Prefetching

- Adaptive and online learning



Rewards:
- Accurate and timely
- Accurate but late
- Loss of coverage
- Inaccurate
- No-prefetch

Bera, Rahul, et al. "Pythia: A customizable hardware prefetching framework using online reinforcement learning." *MICRO-54,* 2021.

Chip Design and its Impact on AI

Why Machine Learning or AI is so Popular

- Is machine learning a recent phenomenon?

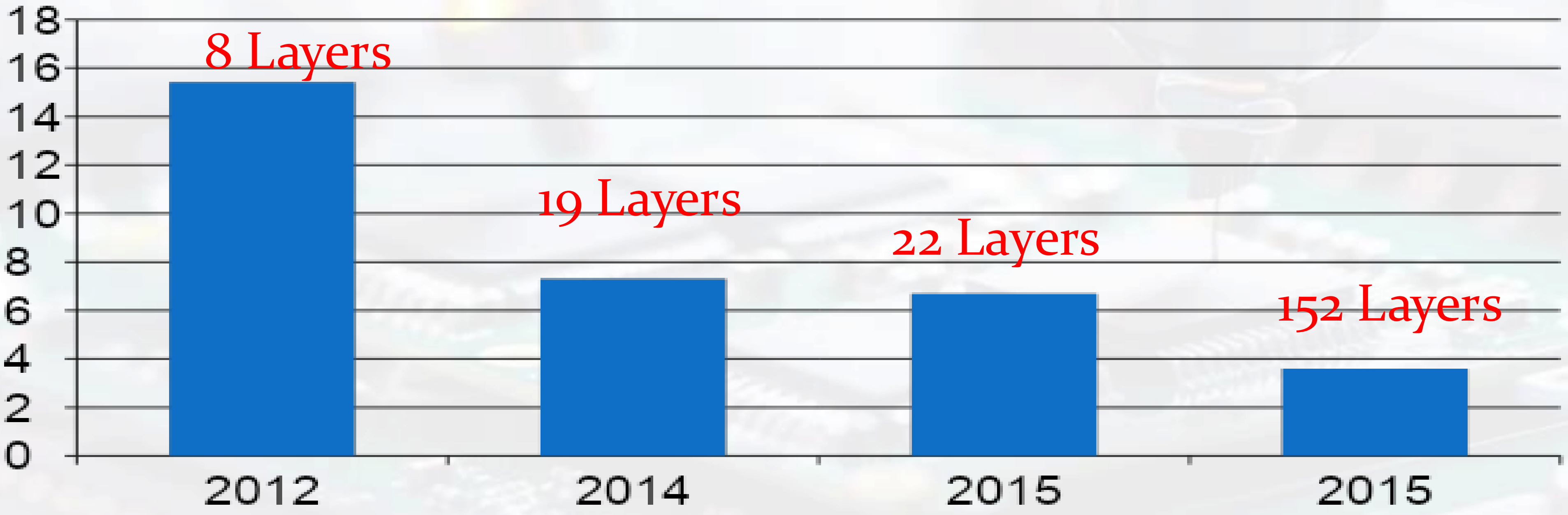
  Samuel AL. Some studies in machine learning using the game of checkers. IBM Journal of research and development. 1959 Jul;3(3):210-29.

- Why so popular now?
  - Abundance of labelled data
  - Abundance of compute and storage power

Evolution of Deep Neural Network

Error rates in ImageNet Challenge over years

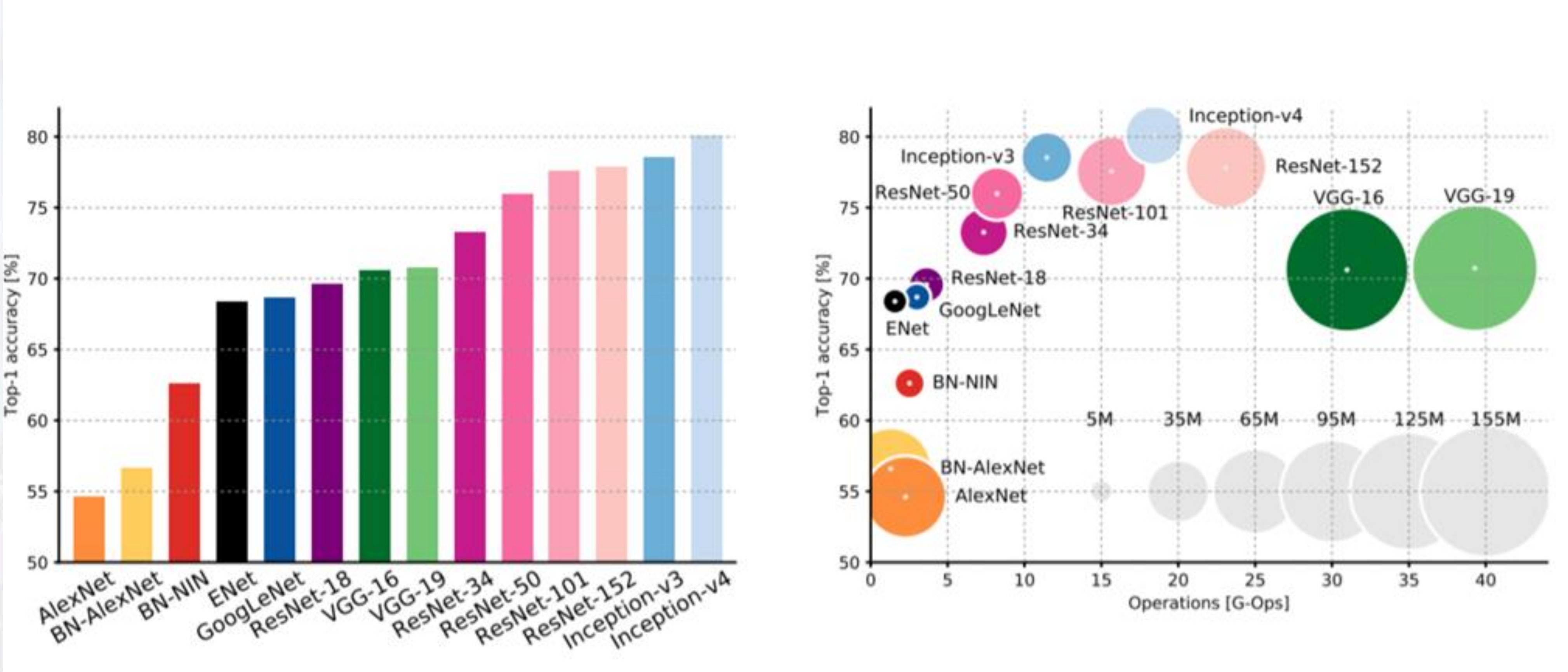
Top 5 Error Rate



AlexNet | VGG | GoogleNet | ResNet

In 2017 ImageNet challenge 29 out of 38 teams had less than 5% error

DNN Architectures



Canziani A, Paszke A, Culurciello E. An analysis of deep neural network models for practical applications. arXiv preprint arXiv:1605.07678. 2016 May 24.

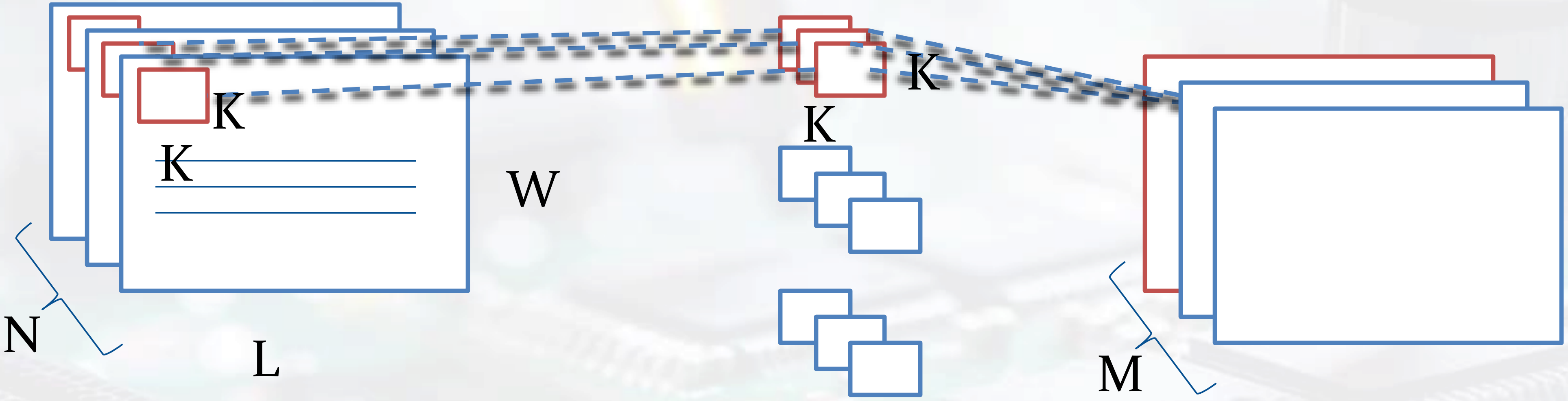
Understanding Computation Requirements of DNN

- Assume computation requirement: 10 Gig operation
- One operation takes 10 compute cycles
- CPU speed 2.5 GHz
- Time for one inference:
  - $10 \times 10^{9} \times 10 / (2.5 \times 10^{9}) = 40$ seconds

Required time per inference : << 40 msec

Trivia:
ChatGPT has 175 B parameters
Took 34 days to train
On 1024 A100 GPUs

Internals of DNN Architecture: Convolution Layer



Input feature map: NxLxW

Kerne
l
NxKx
K

Kernel ~~NxKxK~~ MxNxKxK

Output feature map: L'xW'

Output feature map: ~~L'xW'~~ MxL'xW

Observations

- Lot of inherent parallelism
  - Each pixel in output feature map can be computed in parallel
  - Each feature map can be computed in parallel
  - Each dimension of 3D convolution can be computed in parallel

How GPUs could help?

- What is GPU?

- How does GPU works/How it is faster?
  - Identify work to be done by each pixel (called a thread)
  - Thousands of small cores – each work on a pixel
- Synchronization and thread management issues?
  - All threads are same?
  - Single procedure multiple data (SPMD)

SPMD

- SPMD: Single procedure/program multiple data
- Each processing element execute one thread, works on different data depending on thread id.
- Each PE will have their own control flow
  - May finish in different time
  - What get executed where - invisible to programmer
  - Many similar Pes
- Simplified processor/control design

SPMD Example

Sequential program

```
Void matrix_add(....) {
For(i=0; i < N, i++)
   For(j=0; j < N, j++)
       index = i * N + j;
        C[index] = A[index] + B[index];
}
Main() {
    matrix_add(A, B, C, N);
}
```

- SPMD

```
//tid is calculated based on thread ID
Void matrix_add(...) {
//Int tid;
If(tid < N*N)
C[tid] = A[tid] + B[tid]
}

Main() {
  matrix_add<<<thread_size>>>(A, B, C,
   N)
}
```

How GPU architecture helps in machine learning?

- Machine learning frameworks targets GPUs
  - Keras, PyTorch, TensorFlow
- Library provided by GPU vendors
  - Writing efficient GPU programs is difficult
  - cuDNN by nVIDIA for efficient implementation of DNN kernels
- Cloud computation and cloud storage

Possible alternatives to GPU

- Hardware accelerators
- Application specific processors
  - NPU/TPU
- Embedded processors

AI and Embedded/edge computing

AI/ML with GPUs/Servers on cloud

Application running on laptop or mobile device

-Sends data to cloud

-Cloud do the inference

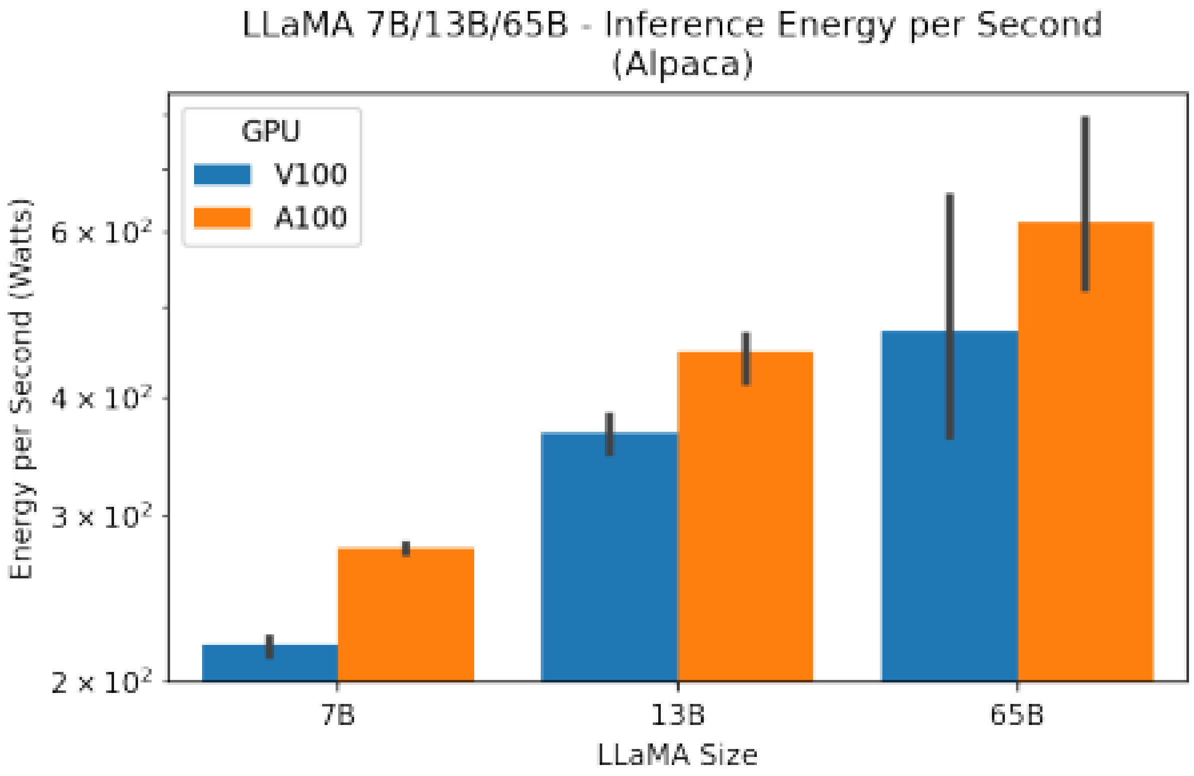
-Sends the results to application on the device

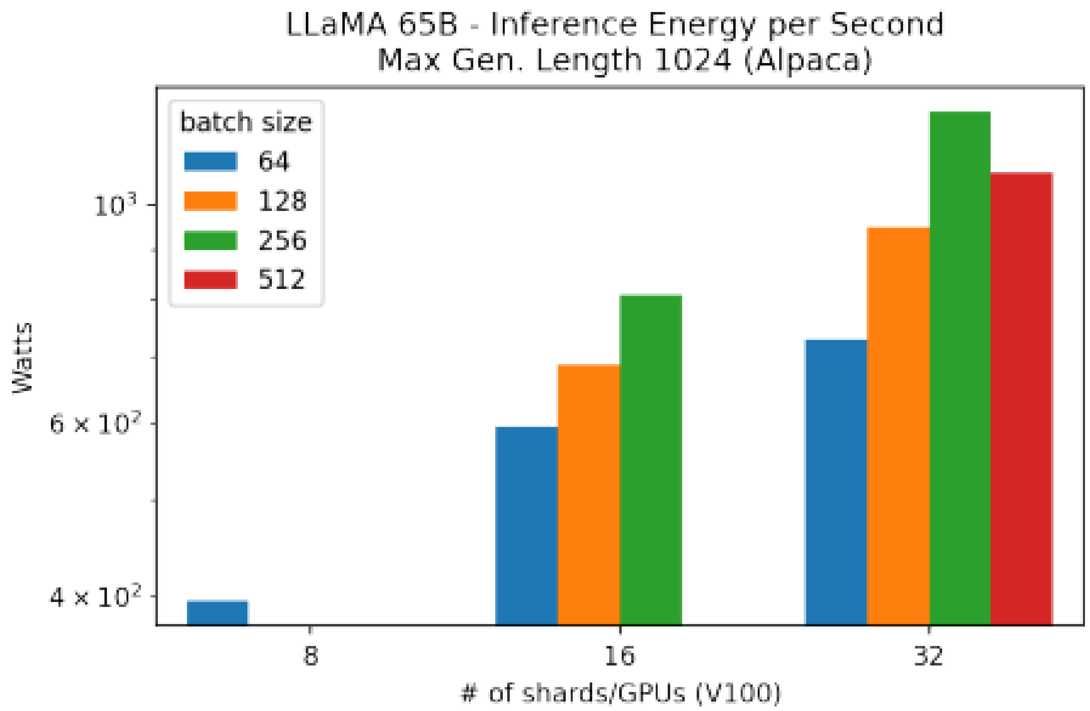
Challenges

-Real time response?

-Cloud cost?

Inference Power Cost at Cloud





Samsi, Siddharth, et al. "From words to watts: Benchmarking the energy costs of large language model inference." *2023 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, 2023.

Can we do ML inference on embedded devices?

Opportunities:

-Multi core devices

-On chip GPU

-NPU accelerator

Example: HiSilicon Kirin 970 SoC

- ARM Cortex-A73 MPCore*4 @up to2.36GHz, ARM Cortex-A53 MPCore*4 @up to1.8GHz
- ARM Mali-G72 MP12 GPU
- 6GB LPDDR4X 1866MHz

Opportunities in Embedded SoCs

- Using GPUs and NPUs
- Exploiting full potential of multiple ARM cores
  - Using threads
  - Using SIMD floating point unit present in each core
  - Use optimized code
- Use all 8 cores together

Potential of optimized ARM code

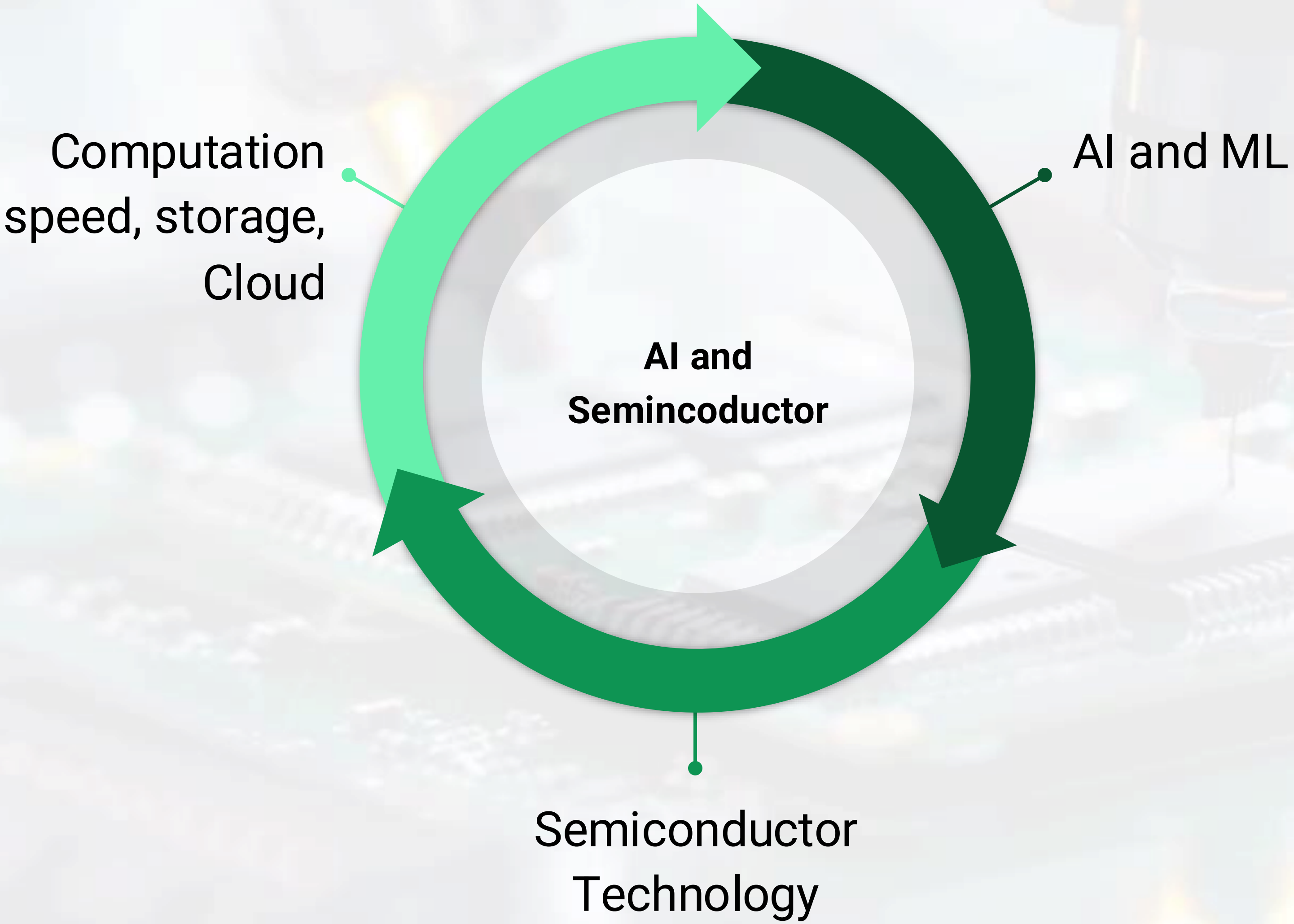
Using ARM Compute Library on HiKey 970 board

Resource utilized	Gaussian 5x5 filter (for processing one image)	Canny edge algorithm (for processing one image)
One A53 core	12.5 msec	77.2 msec
4 A53 cores	4 msec	29.1 msec
One A73 core	6.8 msec	48.37 msec
4 A73 cores	1.98 msec	15.5 msec

Using Cimg library on HiKey 970 board

Resource utilized	Blur (0-order Deriche filter)	Gaussian Blur
One A73	0.48 sec	1.51 sec

Summary


Computation speed, storage, Cloud
AI and ML
AI and Semincoductor
Semiconductor Technology

Q&A

SEMICON® INDIA

ORBIT & SKYLINE

VISIT US AT BOOTH 156

SEP 2-3-4, 2025 | YASHOBHOOMI (IICC), NEW DELHI

Thank You

USA
4930 Campus Drive,
Newport Beach, CA 92660

Sales & Partnerships
hello@orbitskyline.com
om

India
B602, Bestech Business Tower, Sector 66, Mohali
Punjab 160066 INDIA

Job Applicants
careers@orbitskyline.com

HR
hr@orbitskyline.com

Get in Touch
+1-510-509-3202

+91-172-509-9933
9933