Orchestrating a brighter world

NEC

Accelera Taiwan Forum for System Level Verification & Design

AI and IoT Acceleration with FPGA+HLS

Kazutoshi Wakabayashi, Ph.D

NEC Corporation

April 21 2017

Agenda
NEC activities for AI, IoT acceleration

1. Introduction of NEC, myself, HLS(CyberWB)
2. AI algorithm refinement
3. Computing Platform

3.1 Data management

3.2 Heterogeneous computing

3.3 FPGA+HLS acceleration

4. New Devices for AI, IoT

5 Summary

56pages…. quick presentation

1. Self Introduction

1. NEC: ***used to*** be TOP semiconductor company
   - invested a lot to design tools.
   -> ***CyberWorkBench*** was a pioneering HLS
   (why only NEC's HLS succeeded? true?)

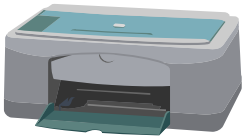
2. Stanford, Yahoo! (founder was HLS researcher)
   internship Jerry Yang: Fujitsu, David Filo: NEC (Kawasaki)

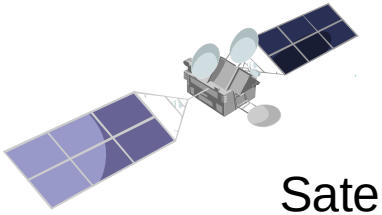
3. Currently, two National Projects:
   "new" **non-volatile FPGA+HLS**
   for accelerating AI and IoT

How to accelerate AI, IoT, etc with low power.

C-based HLS : Many real successes, but often said as "Myth" or …

MFP





Satellite

Cameras

C-based Synthesis

CyberWorkBench

High-end servers

Digital Appliances

Mobile phones

Robots

Automobiles

Reality: some RTL designers cannot not fit HLS, unfortunately.

Our recent activities with “FPGA + HLS”

Traditional areas: Transmission, Graphics and Movies,
Gigabit Ether Phy
4G Base station,
HEVC H.265 - Video CODEC
various graphycal filters, e.g HDR,
JPEG200 like graphic compression
Face recognition, Human extraction, ,,,

- **HFT**(High Frequency Trading) automatic stock trading

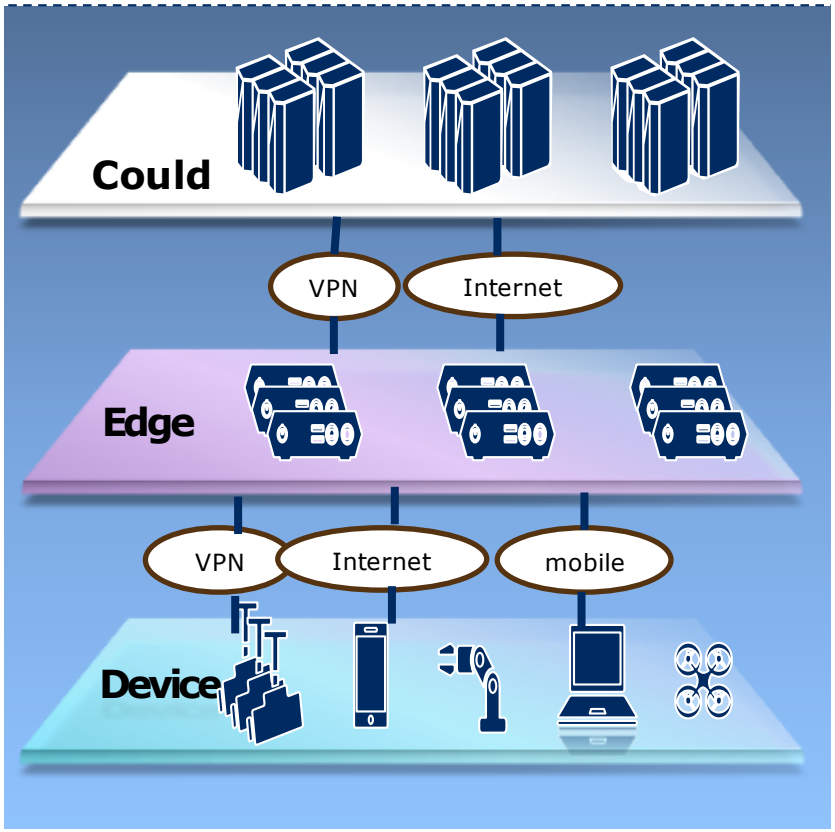
Hot Topics :
- **NFV (Network Functional Virtualization)**
  Networking with IA server (DAC 2016)
  “Xeon” : 10G fine, but 40G, 100G might be tough.

- **AI inference Today’s focus**
  low power, real time for edge device/ computing
  e.g. ADAS, FA,

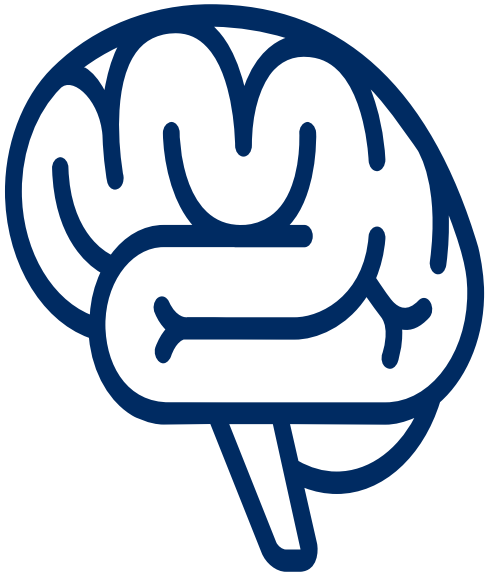
2. Our activities on FPGA

FPGA deployments on our business and research:

IoT Infrastructure

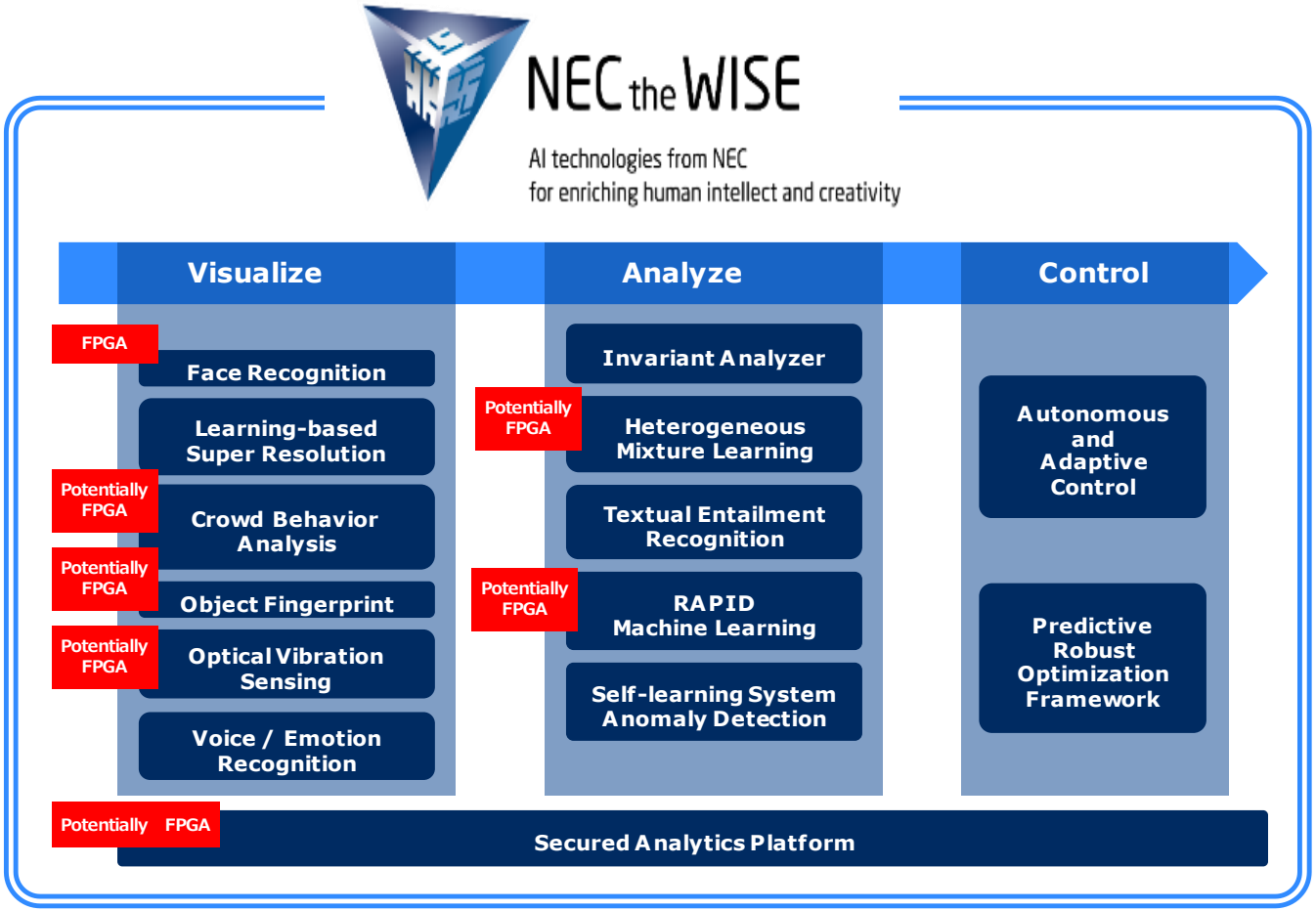


Artificial Intelligence (AI)



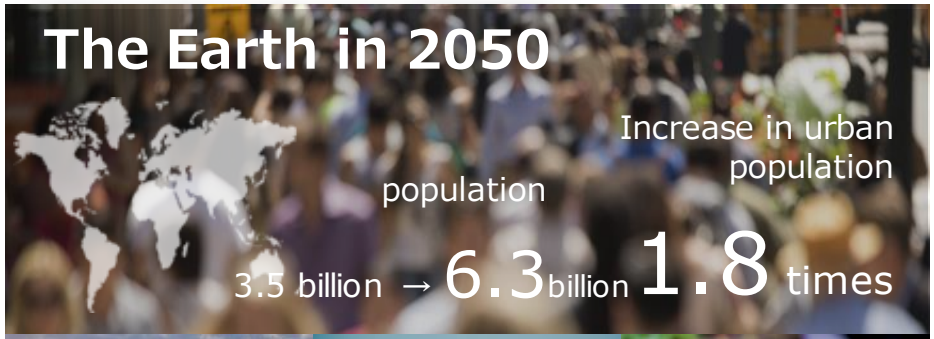
AI: FPGA's opportunities for our AI assets

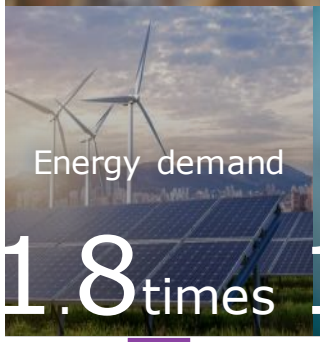
AI-algorithms to be boosted by FPGA

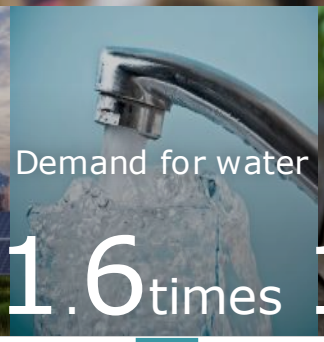


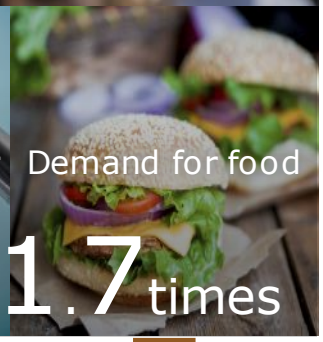
NEC's Social Value Creation

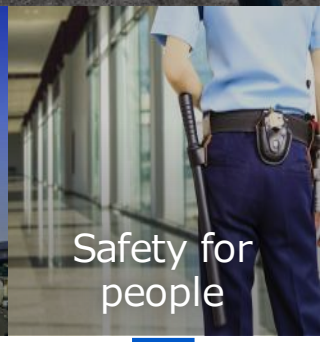
NEC is currently more focusing on the social value creation to solve social issues











Smart energy

Smart water management

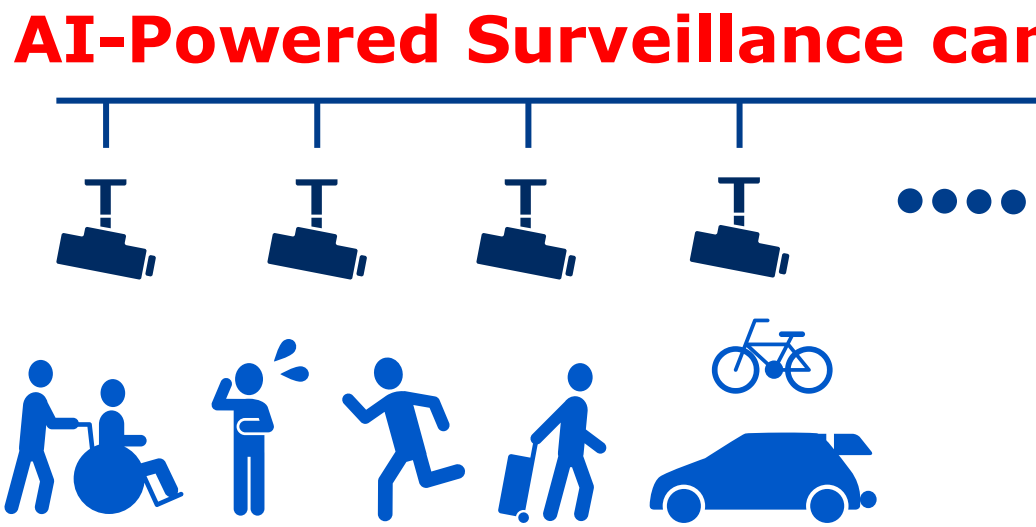
Agricultural ICT

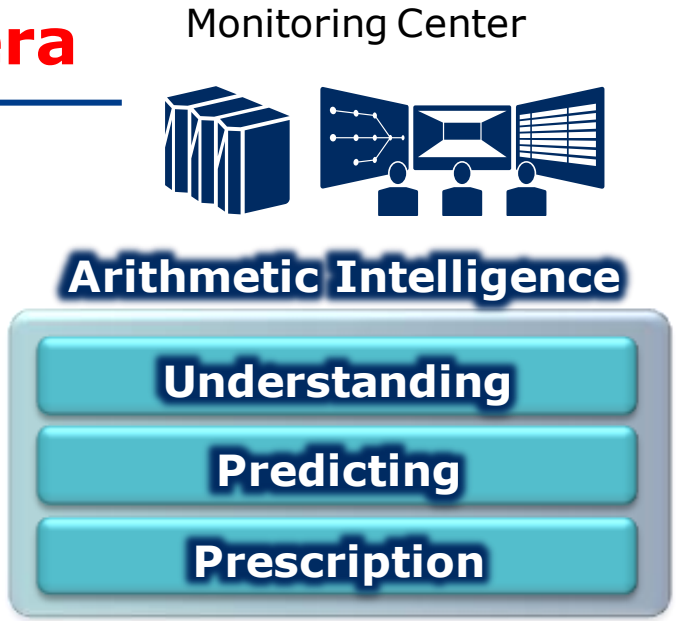
Solutions for enhancing operational efficiency

Diagnosis of infrastructure deterioration

Public safety

Ex. Surveillance system empowered by AI





Need Acceleration! CPU is not fast enough..

Real-time operation for AI applicaiton

"What was bad" -> " Prevent Bad things"

Analyzing in 1 **hour**

Specify Criminal
We could capture him.

Current: Visualization

60x faster

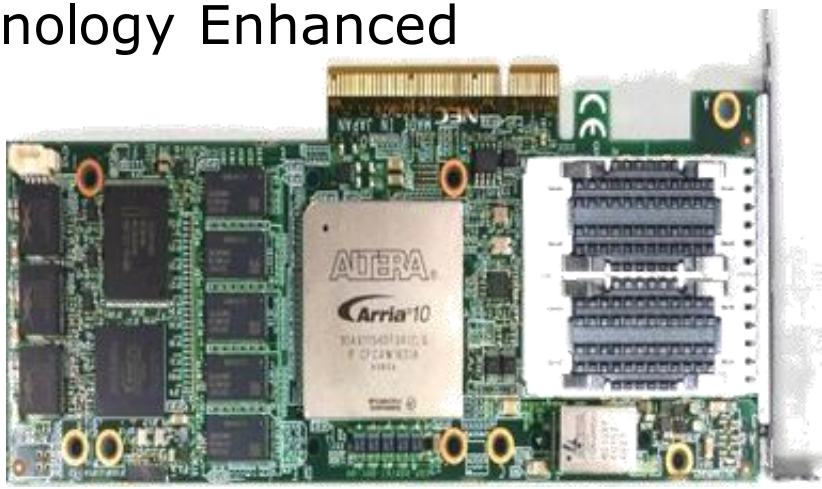
Analyzing within 1 **min**

Specify the sign of crime
We can prevent crime

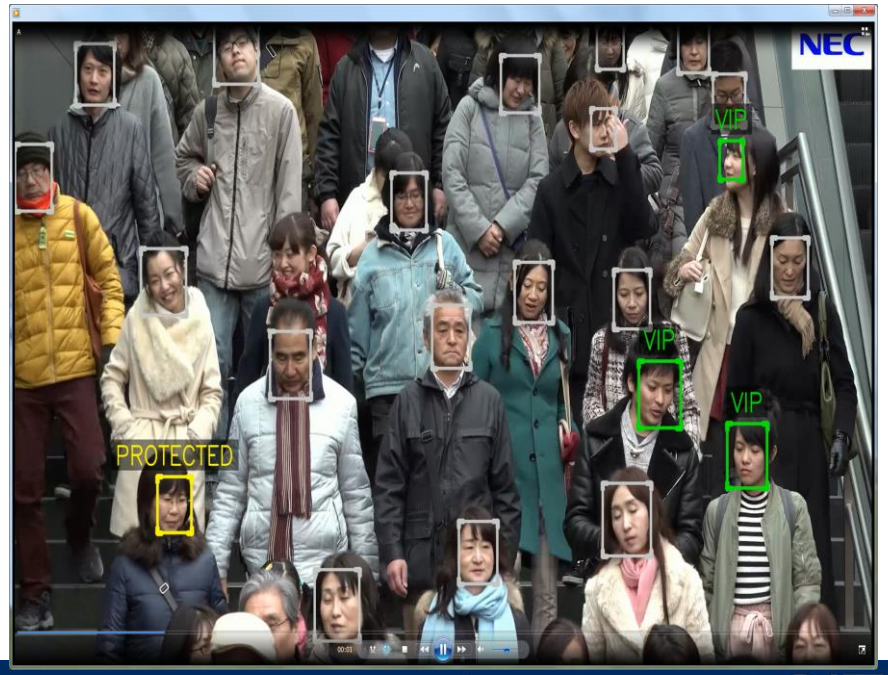
Prevention(Control)

AI: "NeoFace" Acceleration by FPGA Card

- NEC NeoFace Facial Recognition Technology Enhanced by NEC FPGA Acceleration Card



- DEMO: Walk-through Authentication
  - 4K Camera Video Stream
  - Multiple Face Authentication
    20 times faster by the FPGA Card



<https://mipkm.zpf.nec.co.jp/docnavi/DocDetail.aspx?id=1480122>

How to accelerate AI applications with FPGA

refine AI algorithm, and then, tune for FPGA, ASIC

AI algorithms ✖ **Computing platform**

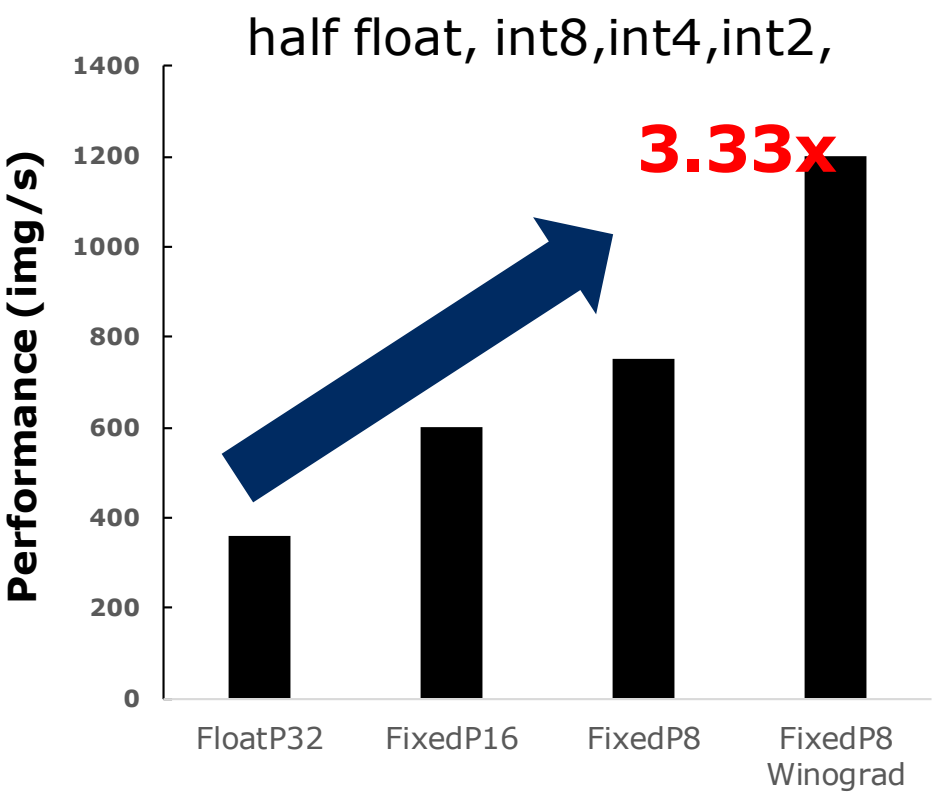
Targets

- **Descriptive Analysis: what is happenig?**
- **Predictive Analisys: What wil happen?**
- **Prescriptive Analysis: What should we do?**

AI: Recent Famous AI algorithm refinement for ASIC/FPGA

precision reduction : higher performance, and better inference

Effects of bit width reduction for CNN



Binarized CNN (**BNN**)

p[i] * input[i]

-> * replace by Xor (^)

p[i] ^ input[i]

FPGA, ASIC
* : 1K,10K,100K,,,
^ : 100K, 1M, 10M,,,

100x, 1000x more

GPU, HPC also aim at float16.int8

How to accelerate AI applications with FPGA

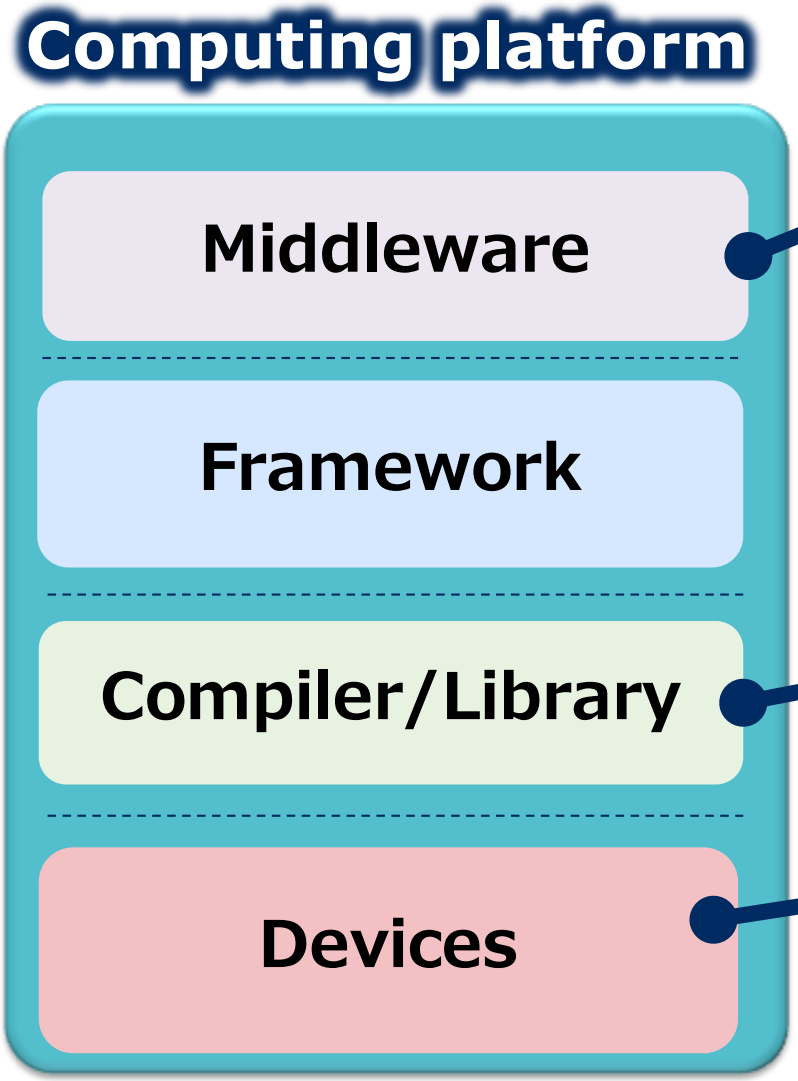
refine AI algorithm, and then, tune for FPGA, ASIC

AI algorithms ✖ **Computing platform**

Targets

- **Descriptive Analysis: what is happenig?**
- **Predictive Analisys: What wil happen?**
- **Prescriptive Analysis: What should we do?**

Our activities for the computing platform for AI



(1) Data management
- Data clustering in D/B

(2) Heterogeneous computing
design tools, libraries

(3) New FPGA
Brain type device

Profiling across spatio-temporal data

analysis of **appearance patterns** of people in video taken for **long-term** and at **multiple points**

AI algorithm

Face recognition

×

Computing platform

Data management

Crime Prevention

Discover **a loiterer** searching for a target (sneak thief, car break-in)

Omotenashi (Hospitality)

Discover **a stray visitor** wandering up and down the street

Profiling across spatio-temporal data

Ex1. appearance patter for time & place
a same person frequently appeared in different places in a day


station platform A


station platform B


a café outside the station

He is possibly a pickpocket seeking targets

Profiling across spatio-temporal data

Ex.2
the same person frequently appeared on/in different dates/areas

How do you think about it? → **He probably did the fire-raising.**

May 4th

Scene A

May 5th

Scene B

May 6th

Scene C

He is possibly an arsonist

How to realize video profiling?

AI engine can extract faces in each video images taken at different places and/or different dates.

AI algorithms

Face recognition

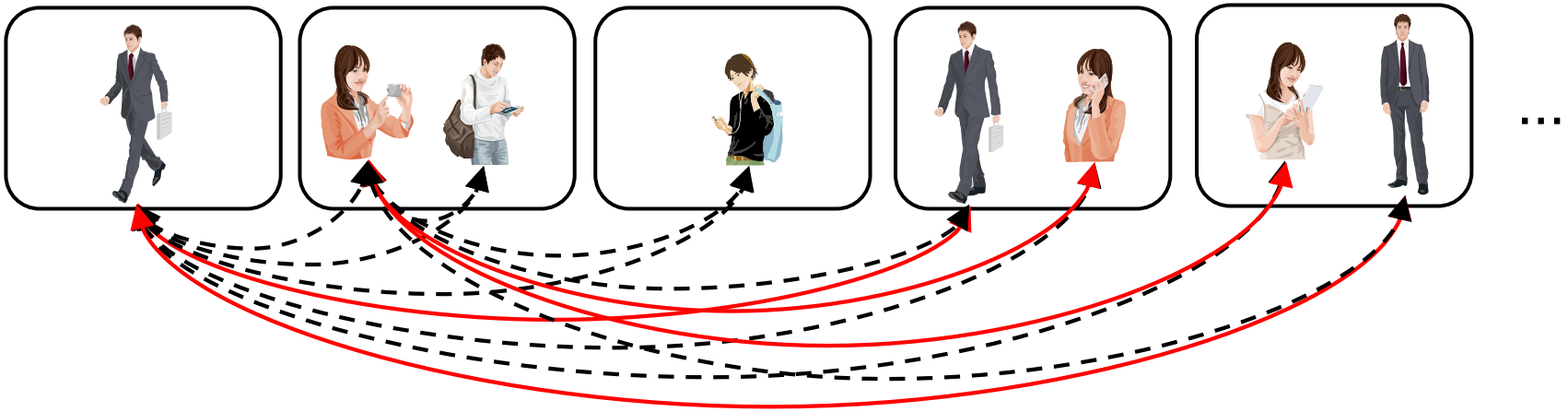
Finding criminal takes long period by simple method

AI algorithms

Face recognition

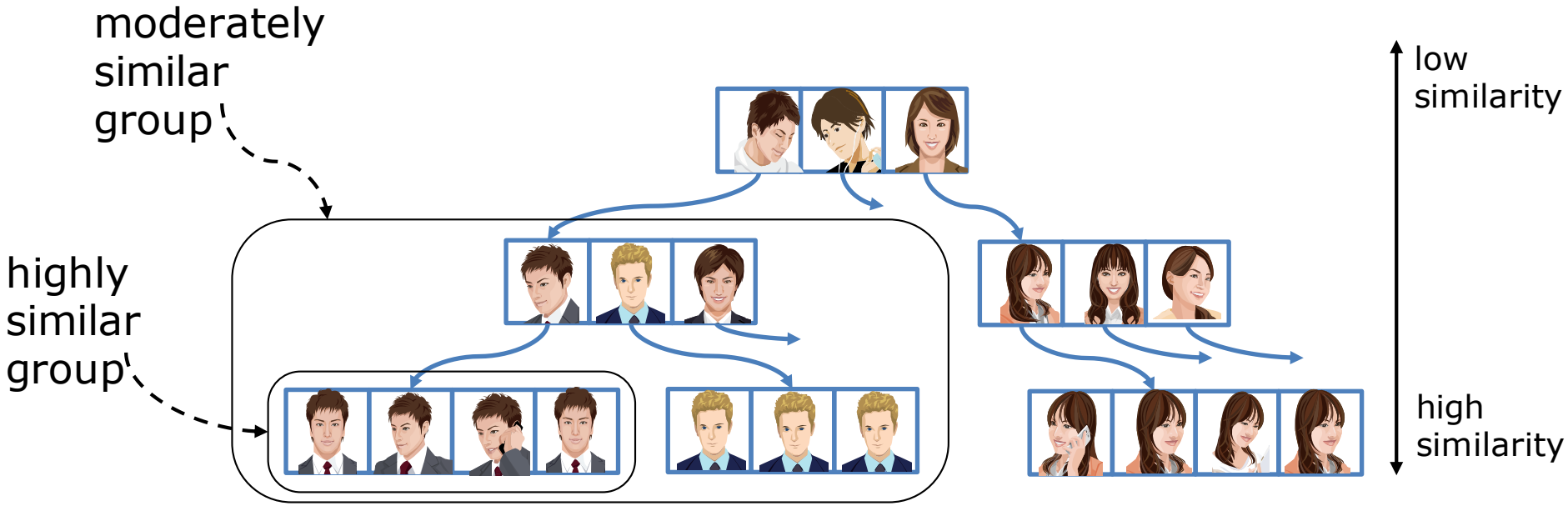
Computing platform

?



Matching one by one takes long computation period.
(Unacceptable)

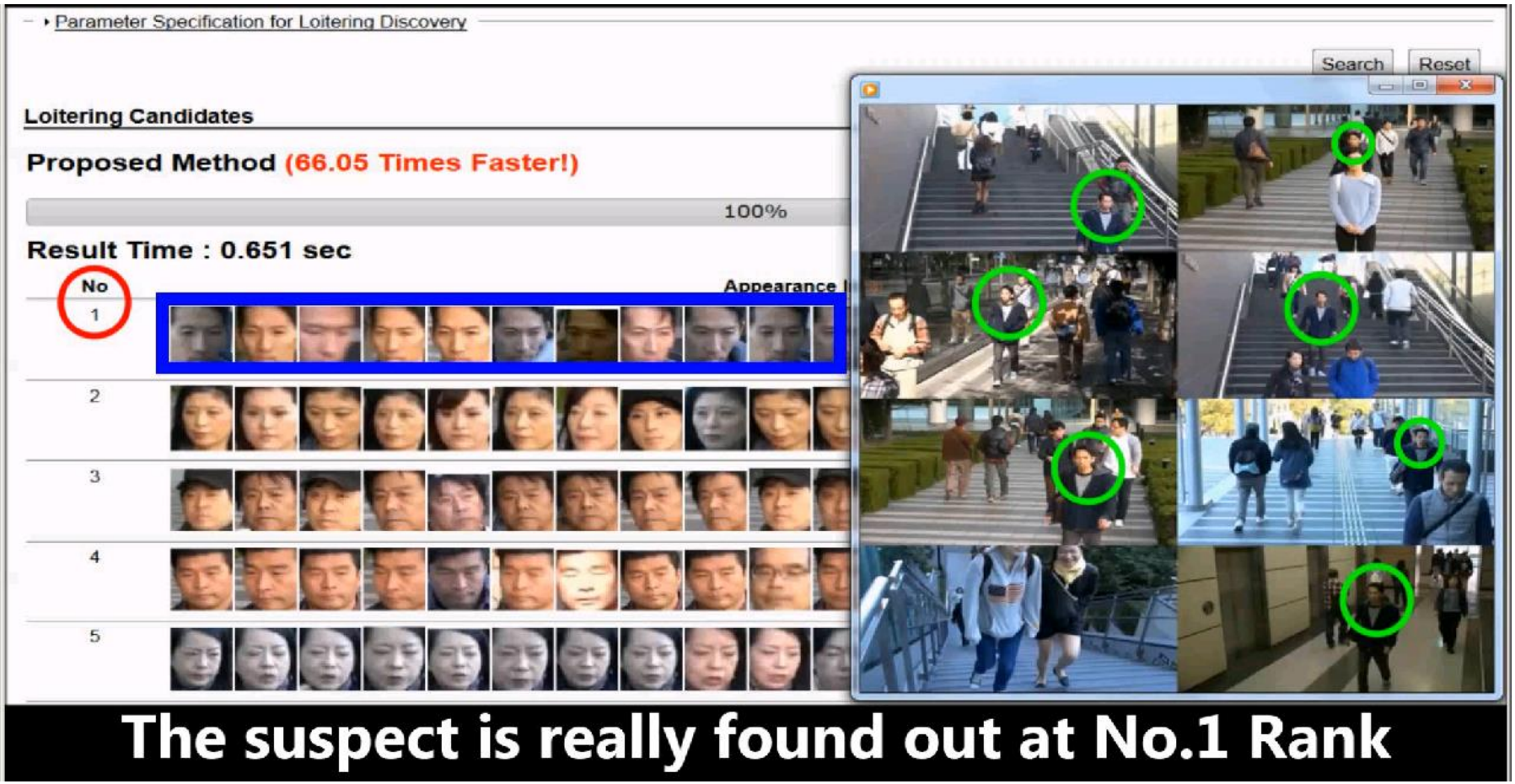
Clustering Faces for a day



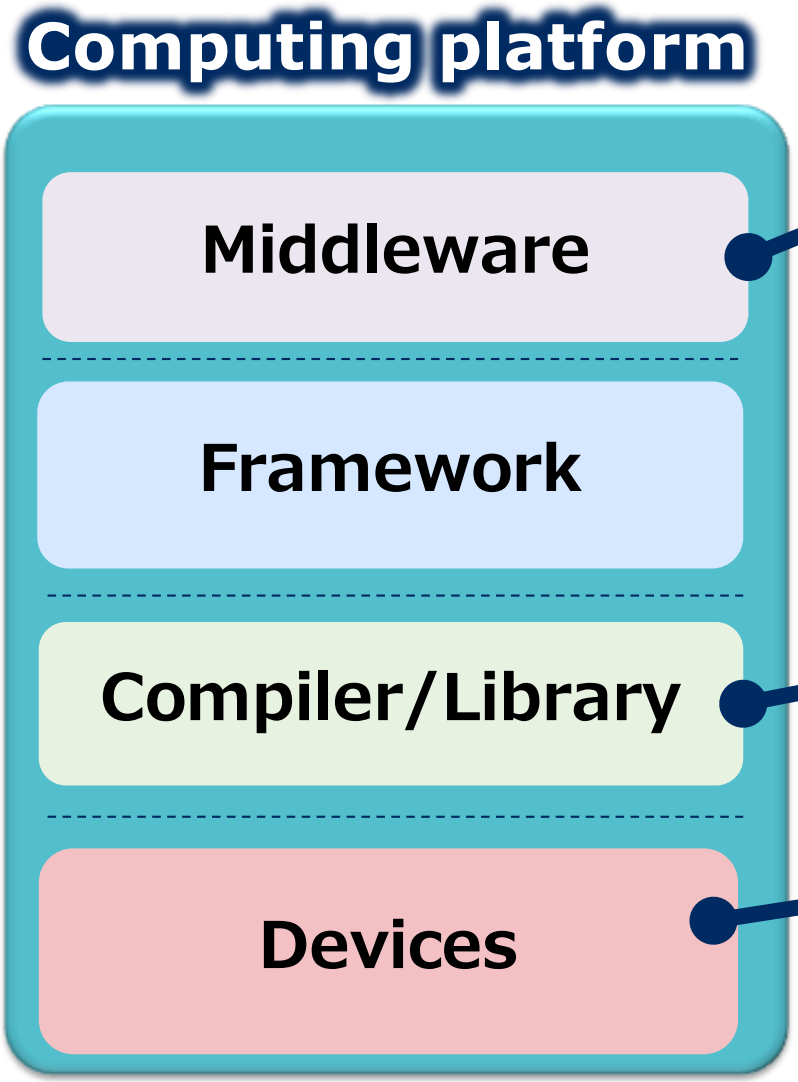
Frequent appeared person will be gathered in a group.
Do not have to match faces .

A representative result

66x faster compared to naïve approach (though not so good….)



Our activities for the computing platform for AI



(1) Data management
- Data clustering in D/B

(2) Heterogeneous computing
design tools, libraries

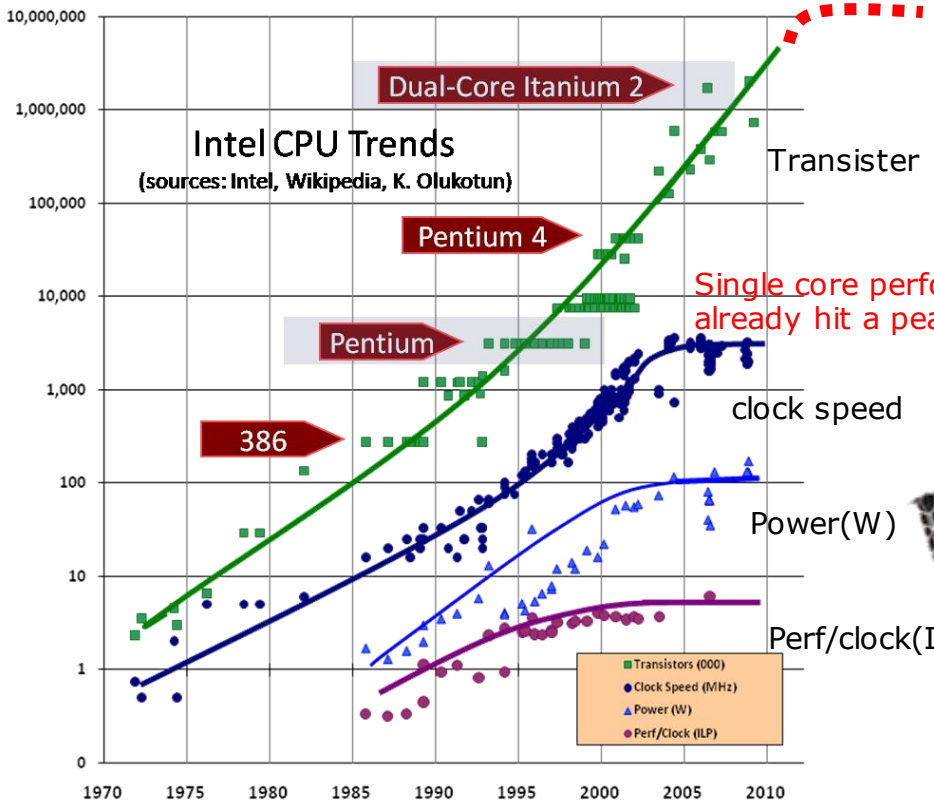
(3) New FPGA
Brain type device

3. Computing Trend

Limit of Moore's law:

more transistors != higher performance

Limit of Moor's low ???? (**Intel deny this)**



http://www.gotw.ca/publications/concurrency-ddj.htm

Single core performance already hit a peak

Single-core(increasing Hz)

Power wall

Multicore/Manycore (increasing # of cores)

Dark Silicon

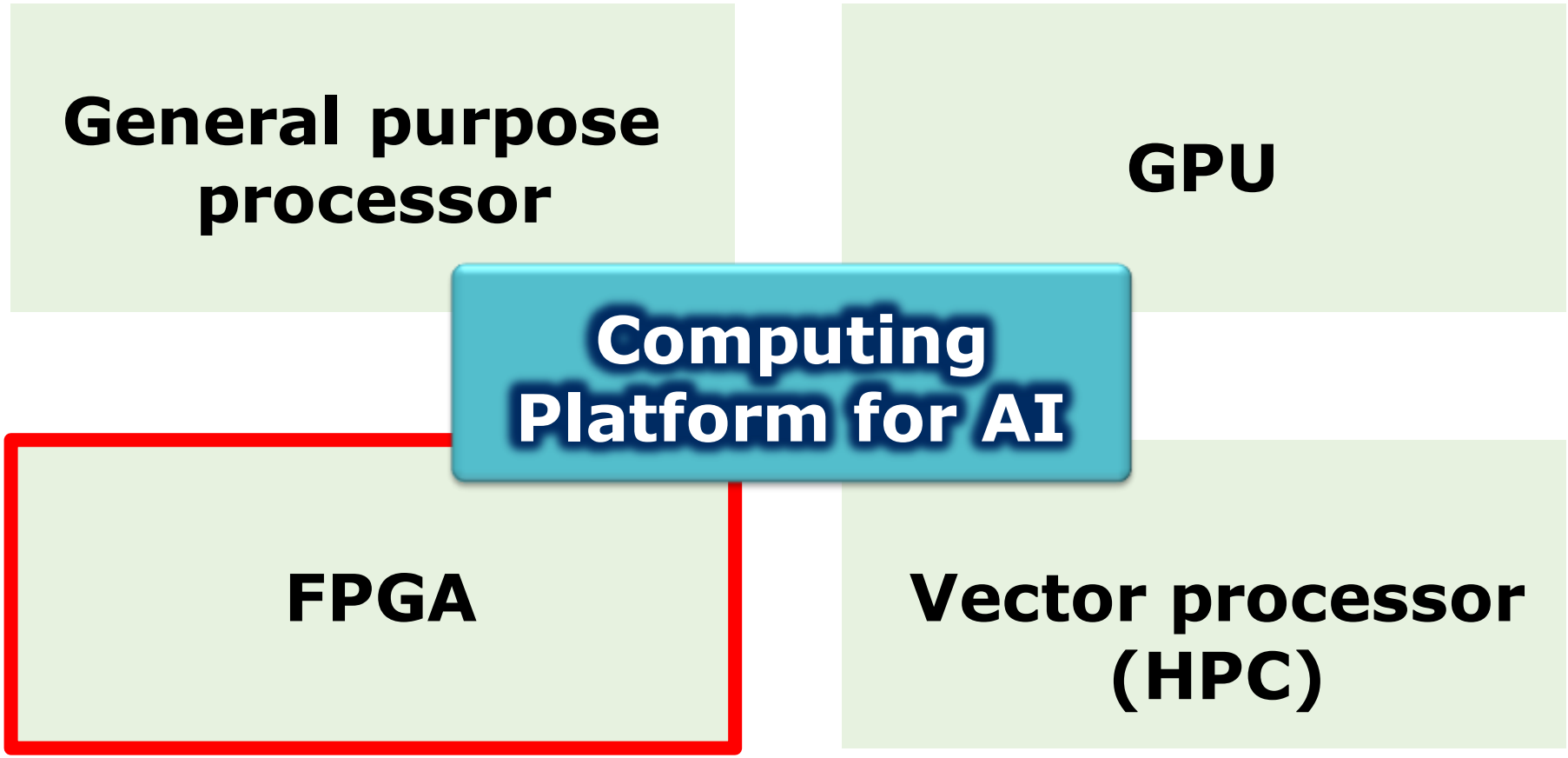
Heterogeneous Computing

http://newsroom.intel.com/docs/DOC-3126

http://www.nvidia.com/object/io_1238654717841.html

Heterogeneous computing

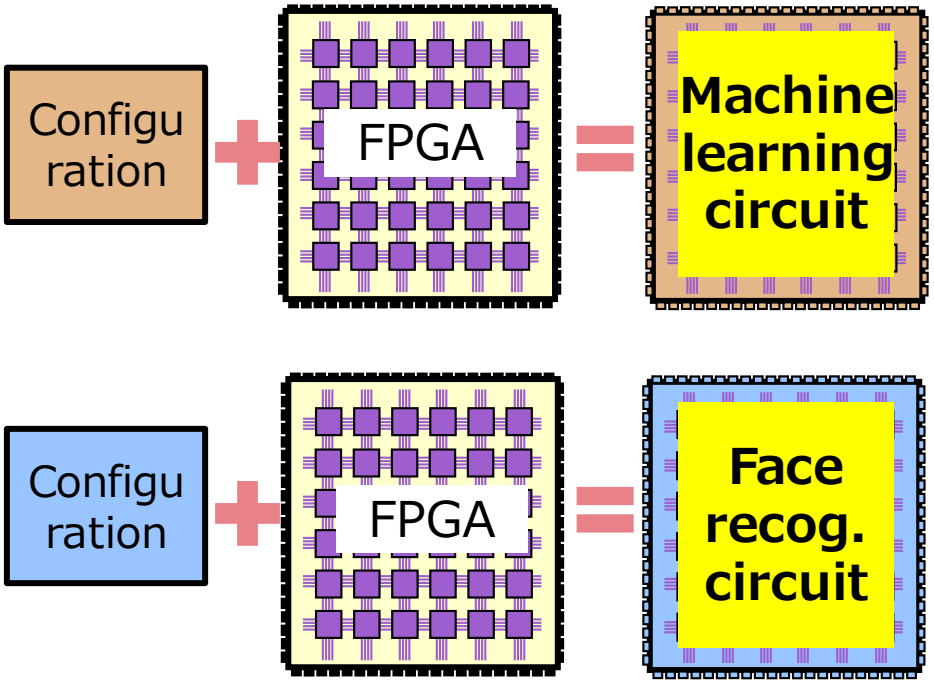
Mixture use of the specialized processors, such as GPUs, FPGAs, and vector processors, **best suited to the application**



What is FPGA

FPGA is a programmable "hardware" for target application

- Enable to construct the dedicated circuits for the application by changing its configuration



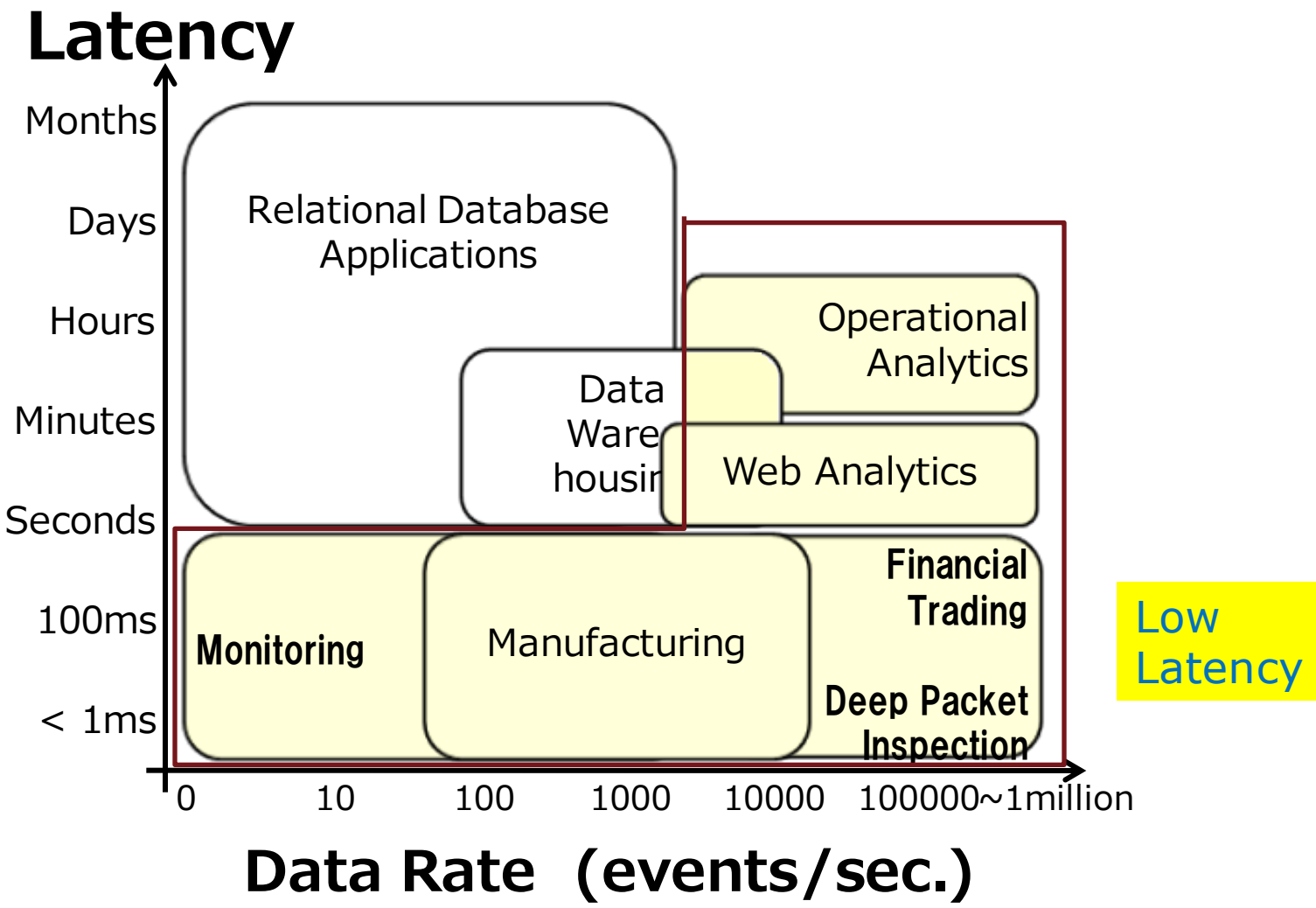
Lower Power
Higher Performance

than Processor

Execution of AI (DNN, CNN, BNN etc.)

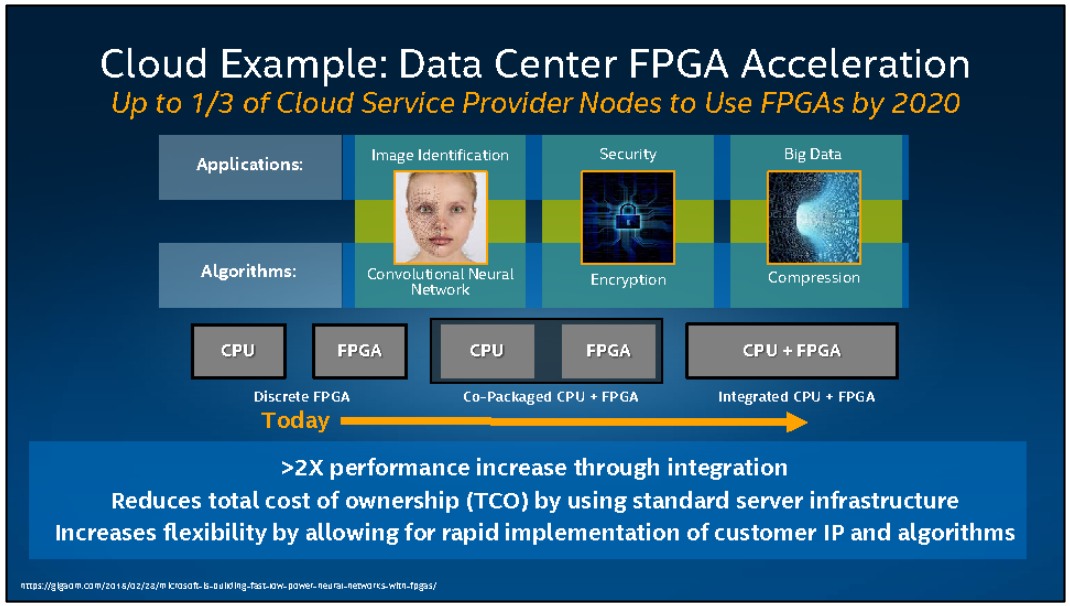
Feature	Type	example
General & slow	CPU	Xeon
	many core	Xeon Phi
	CPU+GPU	Nvidiia+CUDA
	CPU+**FPGA**	**HLS or, RTL**
Special & fast	CPU+**ASIC**	Xeon + Crest

good area for FPGA acceleration



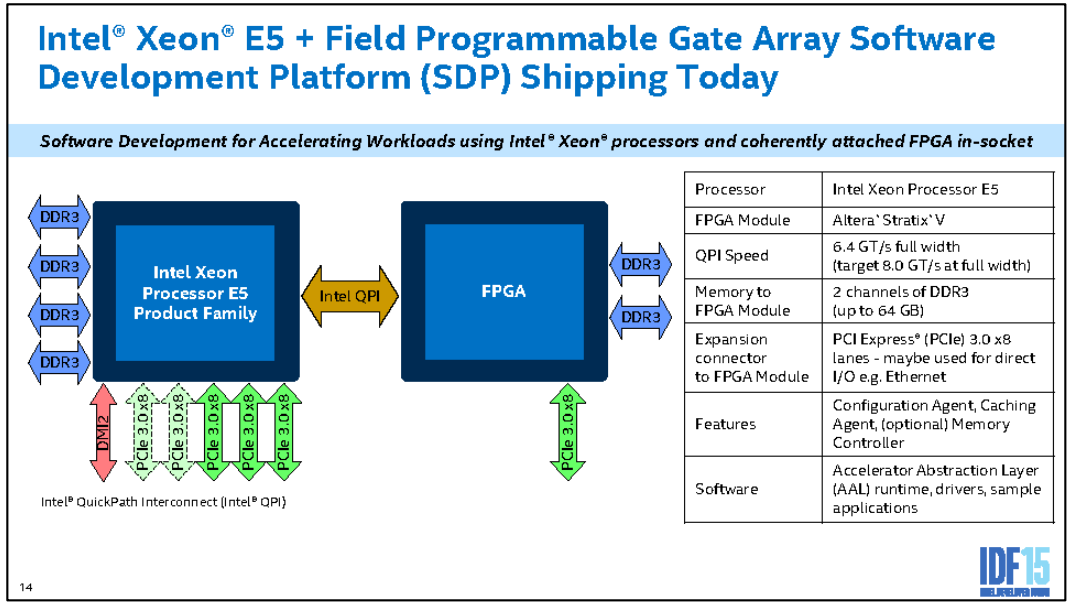
modified : courtesy by M. Patel, et al. "Complex Event Processing: Power your middleware with StreamInsight", Microsoft Tech-ed 2011

FPGA tightly coupled with CPU (FPGA+CPU)



Intel acquired Altera, which is one of the largest FPGA vendors

Intel "Acquisition of Altera" (2015/06/01).



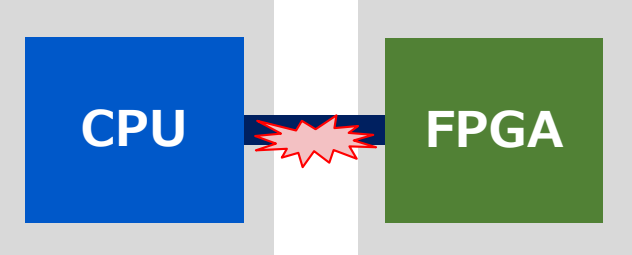
Announced the plan to integrate **FPGA and CPU into a single package**

P. K. Gupta "Using Field Programmable Gate Array to Accelerate Application Performance", IDF 2015 DCWS008.

Benefits of FPGA tightly coupled with CPU

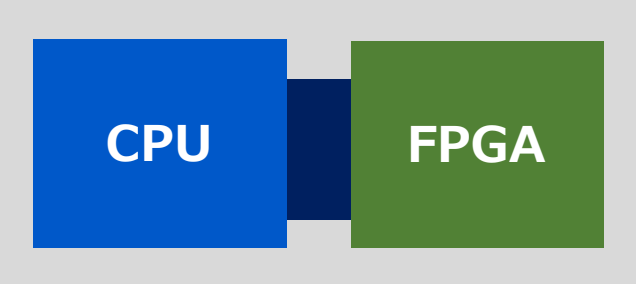
broadband interconnect between FPGA and CPU.

Conventional:
connected with PCIe

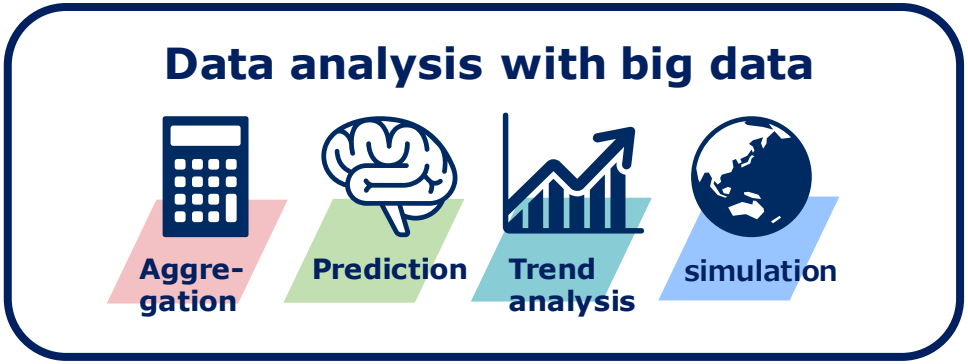


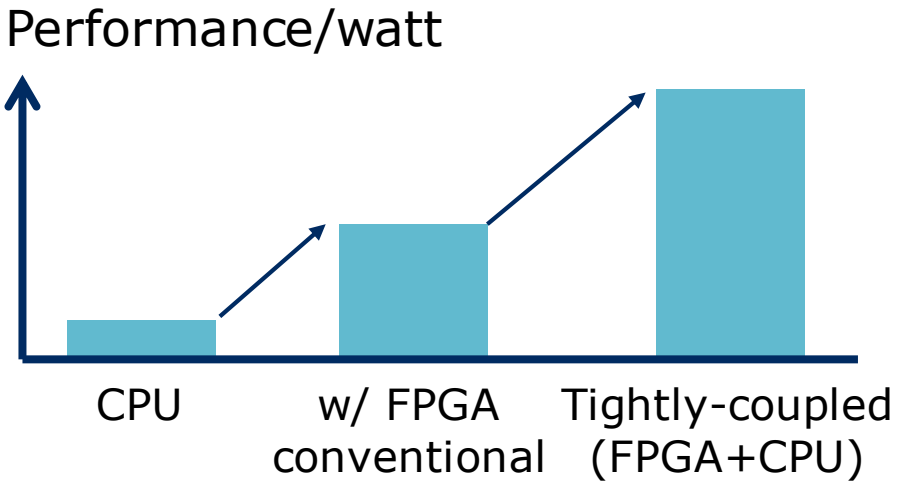
Data transfer is to be the bottleneck

Tightly-coupled (FPGA+CPU)
Wide-bandwidth/low-latency



Enable to accelerate wide range of applications





Source: PK Gupta: "Using a Field Programmable Gate Array to Accelerate Application Performance," IDF'15, DCWS008, 2015.

AI acceleration with tightly coupled CPU+FPGA

1. **use HLS**
   - the latest AI algorithm should be on FPGA quickly.
   - **For AI engineers: RTL is low and not acceptable**

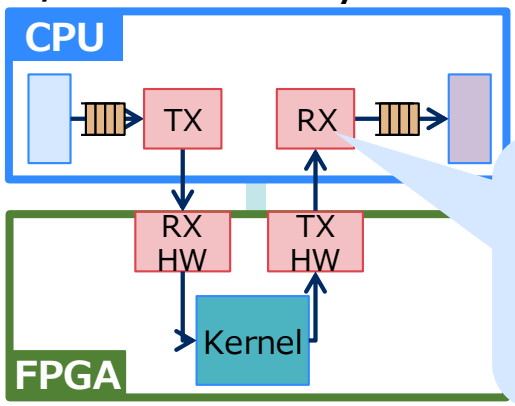
2. **Tune C/C++ program for FPGA/ASIC flexible datapath and control**
   - "FPGA with RTL" and "FPGA with HLS" is not the same

3. fully utilization of **the wideband interconnect** btw. CPU and FPGA

Achievement: CPU and FPGA comm. library

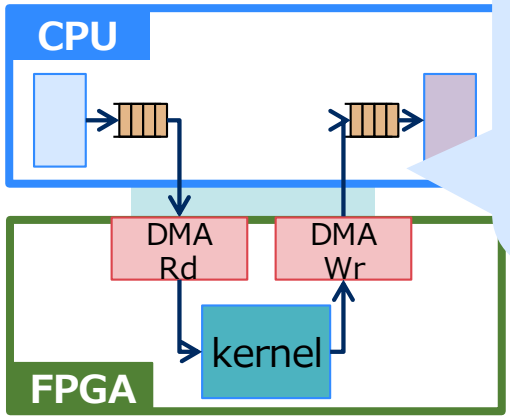
optimized library for broadband interconnect btw. CPU and FPGA.

w/o our library

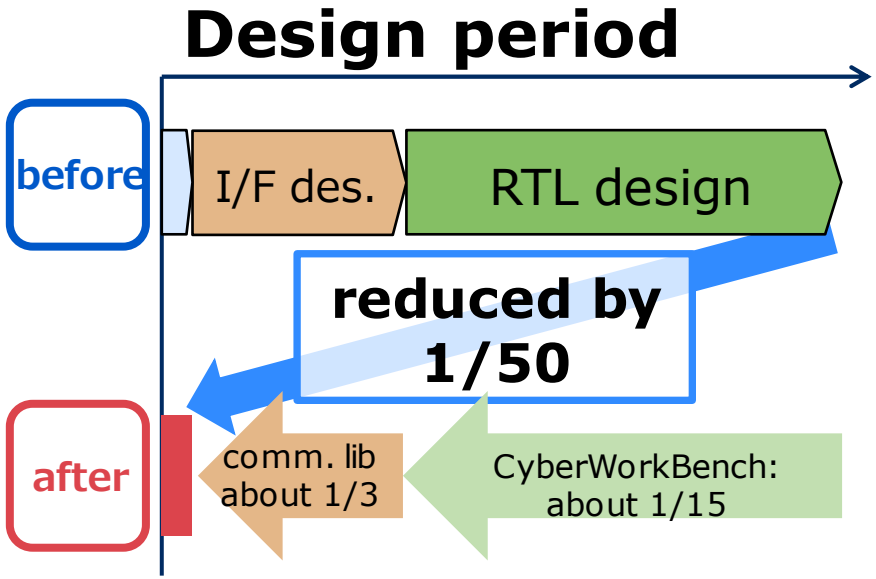


- Need to modify software to communicate with FPGA

w/ our library



- Our library provides optimized communication between CPU and FPGA



※1: Estimated by NEC based on code size.

FPGA+HLS vs CPU/GPGPU + CUDA, OpenCL,,

"FPGA+HLS" is not same as "FPGA+RTL"

1) Architecture View
   - Flexibility of controller and datapath

2) Compiler(HLS) View
   - Control Dependencies
   - Data Dependencies

HLS : target architecture is **Not same as RTL**

C program

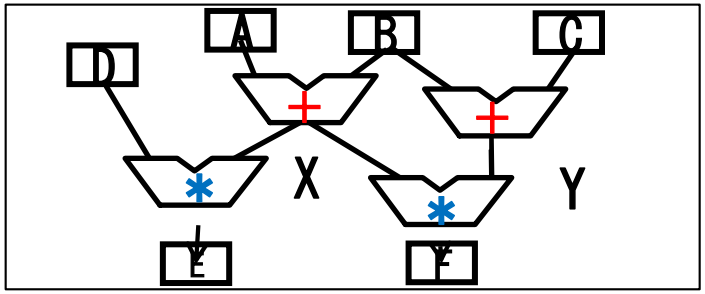
```
char A,B,C,D;
char E,F;
main(){
char X;
X = A + B;
E = X * D;
F = (B + C) * X;
}
```

Clock 5ns

FU constraint

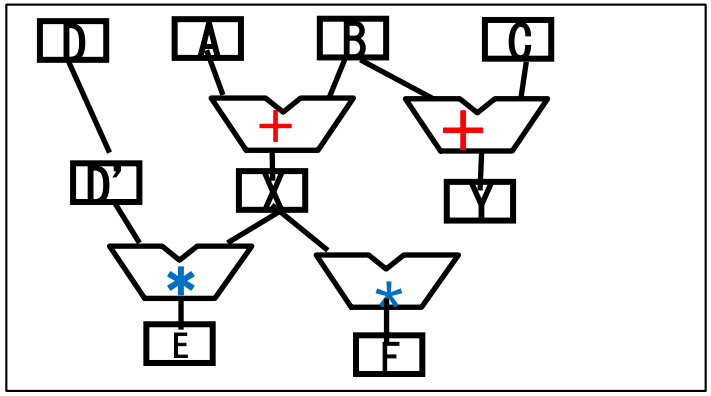
\+ : 2

* : 2



1 cycle

Clock 3ns



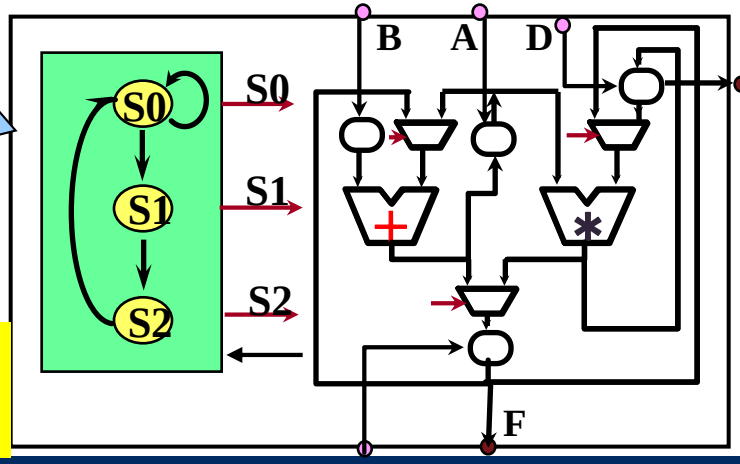
Pepeline
DII=1,
Latency=2

FU constraint

\+ : 1

* : 1

Clock 3ns



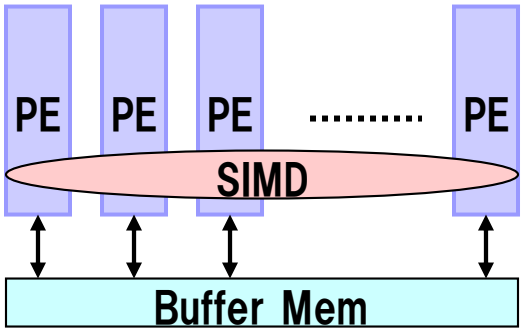
3cycle
Sequential
Circuit

small

RTL designer
does not apply this

SIMD(GPGPU) vs FPGA+HSL: Architecture

SIMD CPU



- Fixed Datapath

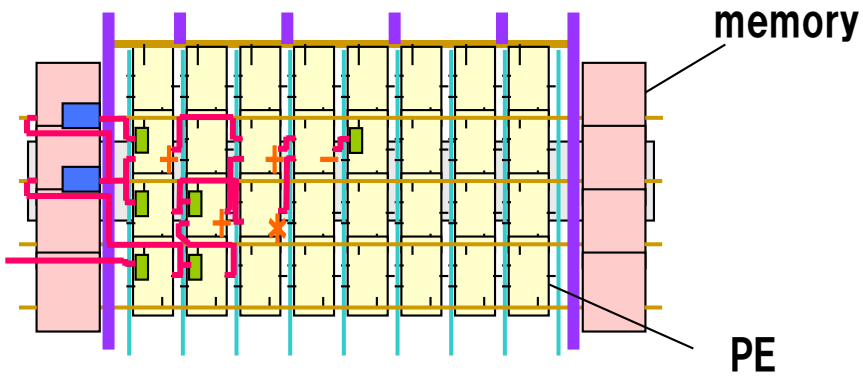
Fixed access among PE and RAM
limited array access.

- Fixed controller:

weak for branches,
complex control

# of core, # of thread is limited.

FPGA/reconfigurable chip



RTL

Flexible DataPath
- flexible and many wires
- flexible access to RAM

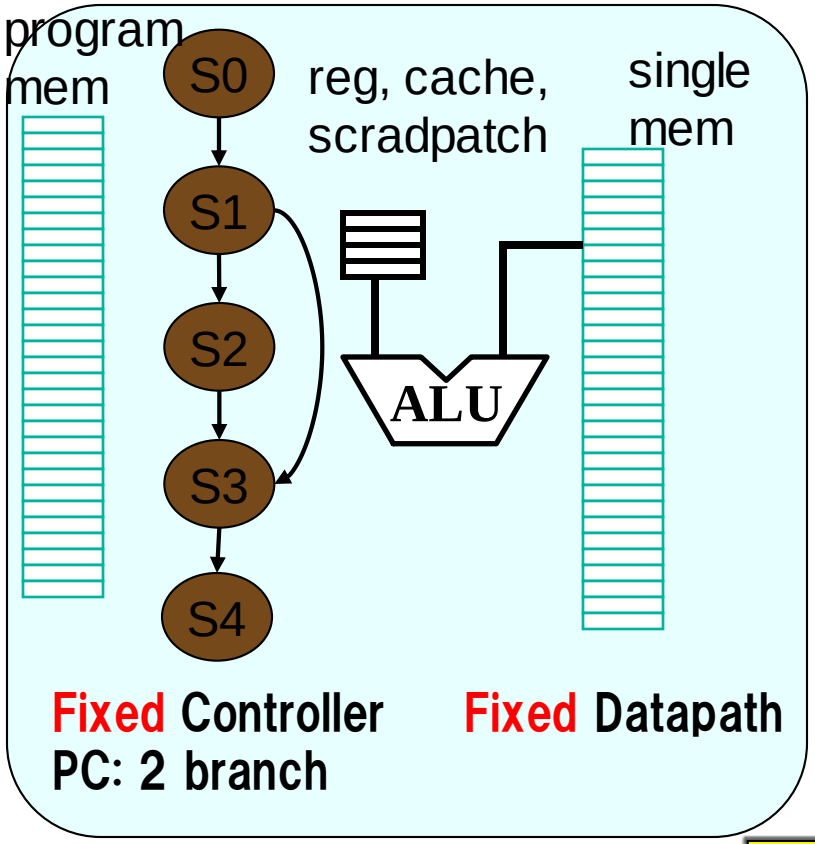
HLS

Flexible controller
- free from control depencency

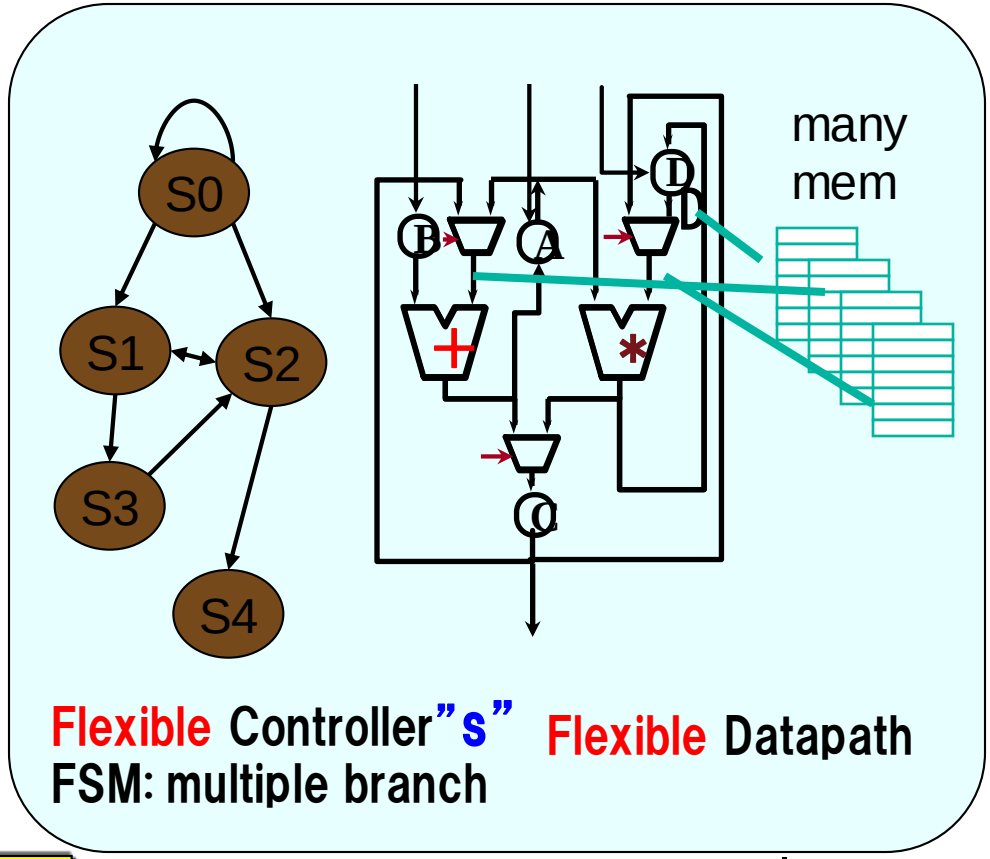
as many as necessary processes
(FSM+Datapath)

Models for CPU and FPGA in “FSM+Datapath” model

model for CPU



model for FPGA + HLS



optimized ALUs: Low Delay
RTOS: task scheduling
Von-Neumann Bottleneck!

GPGPU

SIMD operations

Low Latency , Larger Delay
high parallelism for complex controls (IF, LOOP, function..)
multiple tasks: parallel

“FPGA + HLS” : two important keys for acceleration

FPGA: Flexible controller and datapath
CPU: Fixed controller and datapath

	CPU/GPU	FPGA+RTL	FPGA+HLS
Conditional Branch	**weak**	**tough**	**free from dependency**
Array access	**fixed**	**manual division is tough**	**automatic array division**

Current HW algorithms require decisions

ex. Variable Length code Decoder: (VLD)

```
macroblock()
{
  if (vop_coding_type != "B") {
    if (video_object_layer_shape != "rectangurlar && !(sprite_enable && low_latency_sprite_en able && sprite_tranmit_mode == "update"))
      mb_binary_shape_coding()
    if (video_object_layer_shape != "binary_only") {
      if (!transparent_mb()) {
        if (vop_coding_type != "I" && !(sprite_enable && sprite_transmit_mode == "piece"))
          not_coded
        if (!not_coded || vop_coding_type == "I") {
          mcbpc
         if (!short_video_header && (derived_mb_type == 3 || derive_mb_type == 4))
            ac_pred_flag
          if (derived_mb_type != "stuffing")
            cbpy
          else
            return()
          if (derived_mb_type == 1 || derived_mb_type == 4)
            dquant
          if (interlaced)
            interlaced_information()
          if (!(ref_select_code == '11' && scalablility) && vop_coding_type != "S") {
            if (derived_mb_type == 0 || derived_mb_type == 1) {
              motion_vector("forward")
              if (interlaced && field_prediction)
                motion_vector("forward")
```

1bit code

1～9bit VLC

1bit code

1～6bit VLC

2bit code

VLC Table: I-VOP mcbpc

CODE	MB TYPE	cbpc
1	3	00
001	3	01
010	3	10
011	3	11
0001	4	00
0000 01	4	01
0000 10	4	10
0000 11	4	11
0000 0000 1	Stuffing	—

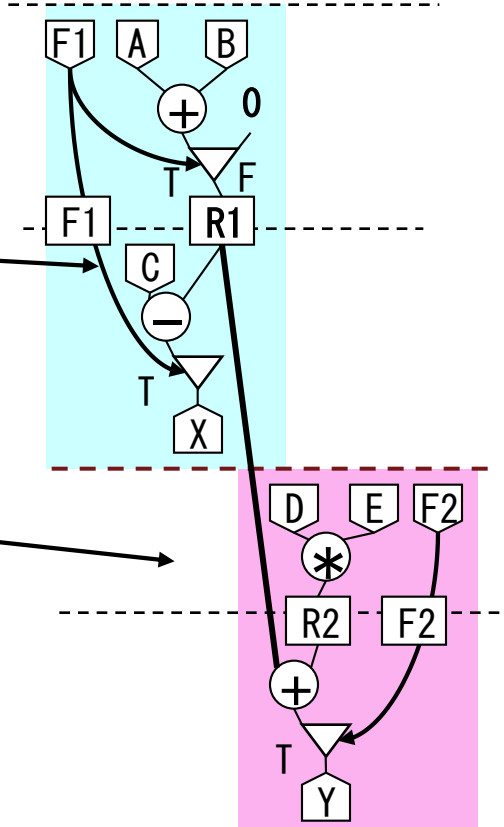
Current Signal Processing contains many decisions and bit handlings

Function: If 25th bit of R3 is 0, then goto L1

PFGA: control dependencies is not barriers for parallelization
Parallelization of Multiple **Branches, loops, functions**

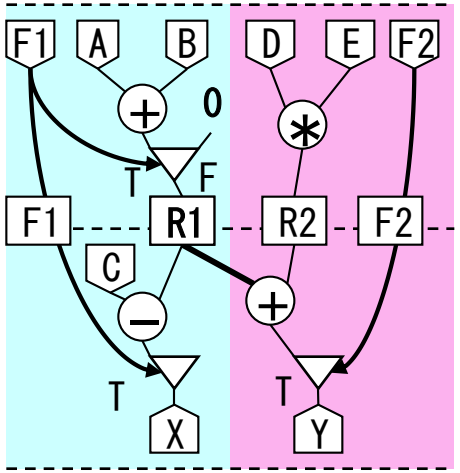
```
if( F1 ) {
    R1 = A + B ;
    X = R1 - C ;
} else  R1 = 0 ;

if( F2 ) {
    R2 = D * E ;
    Y = R1 + R2 ;
}
```

CPU S/W compiler



4 cycles (max.)

FPGA H/W compiler

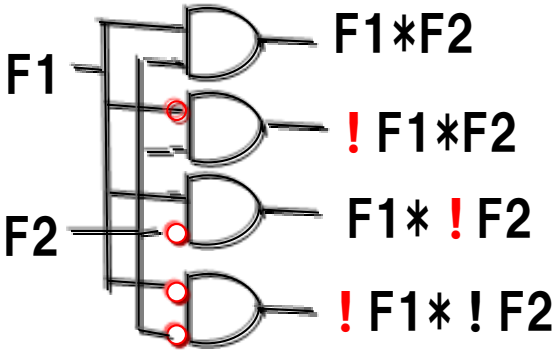


2 cycles (max)

can parallelize several branches

branch unit can be create with AND gate

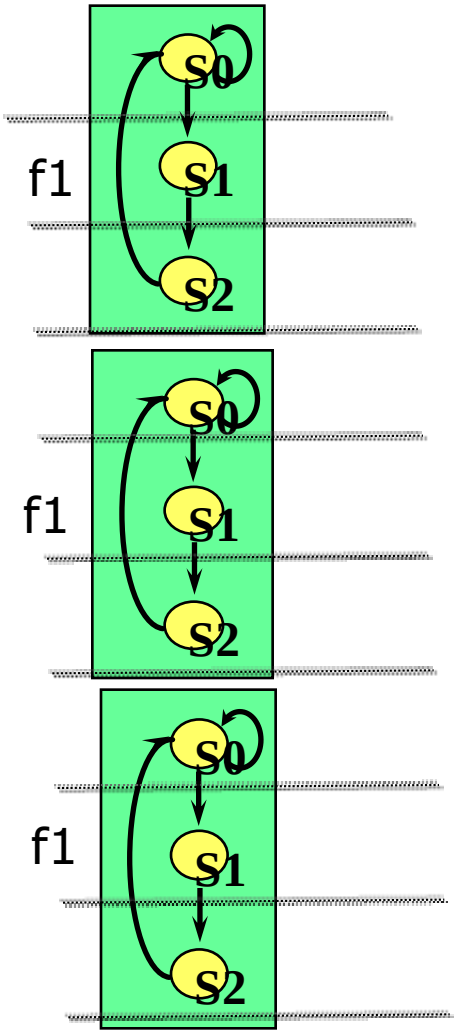
branch condition signals

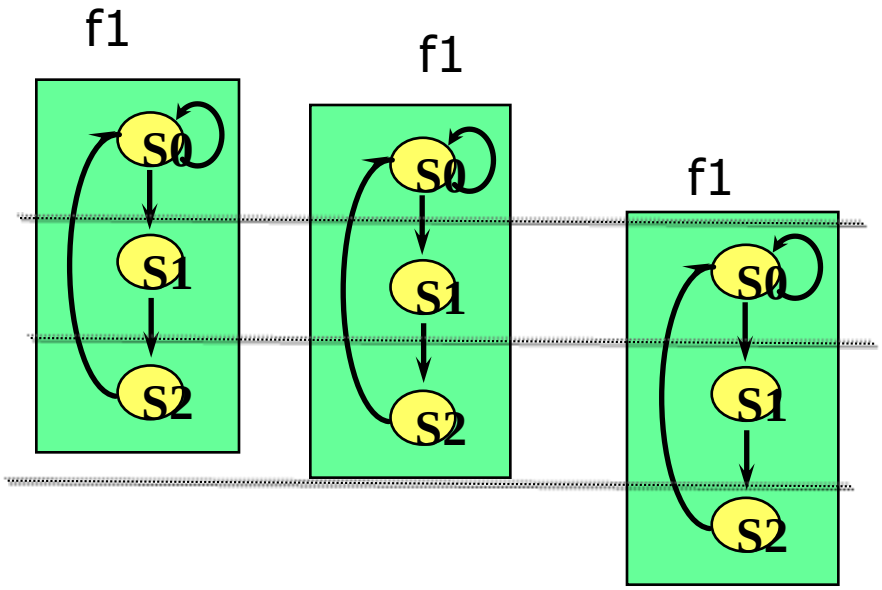


CPU
# of branch unit is fixed

Function Instantiation as many as necessary

```
Function f1()


f1();
if ..
  f1();
else
  ..
  f1();
a++;
....
f1();
```





CPU:
Single Program Counter

FPGA+HLS
functions can be parallelize

Map array to multiple RAMs

Ex. Matrix multiply and add

sum[70] = W[500][70] * in[500] + b;

How to parallelize using 200 multipliers

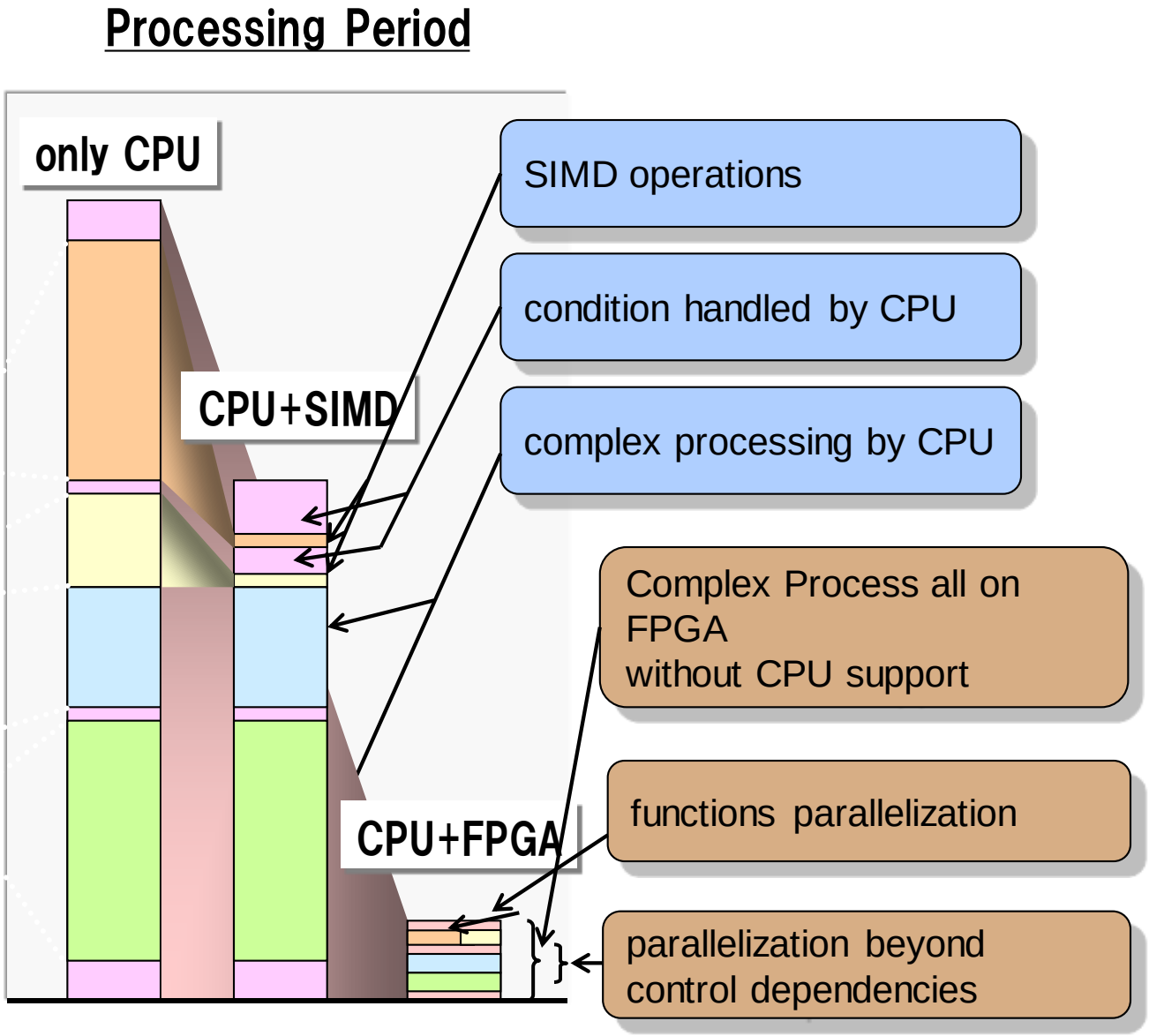
Partition W[500][70] into 200 arrays
into W00[][], W01[][], W02[][], ,,,W10[][],,,

FPGA : many parallel RAM access can be used.

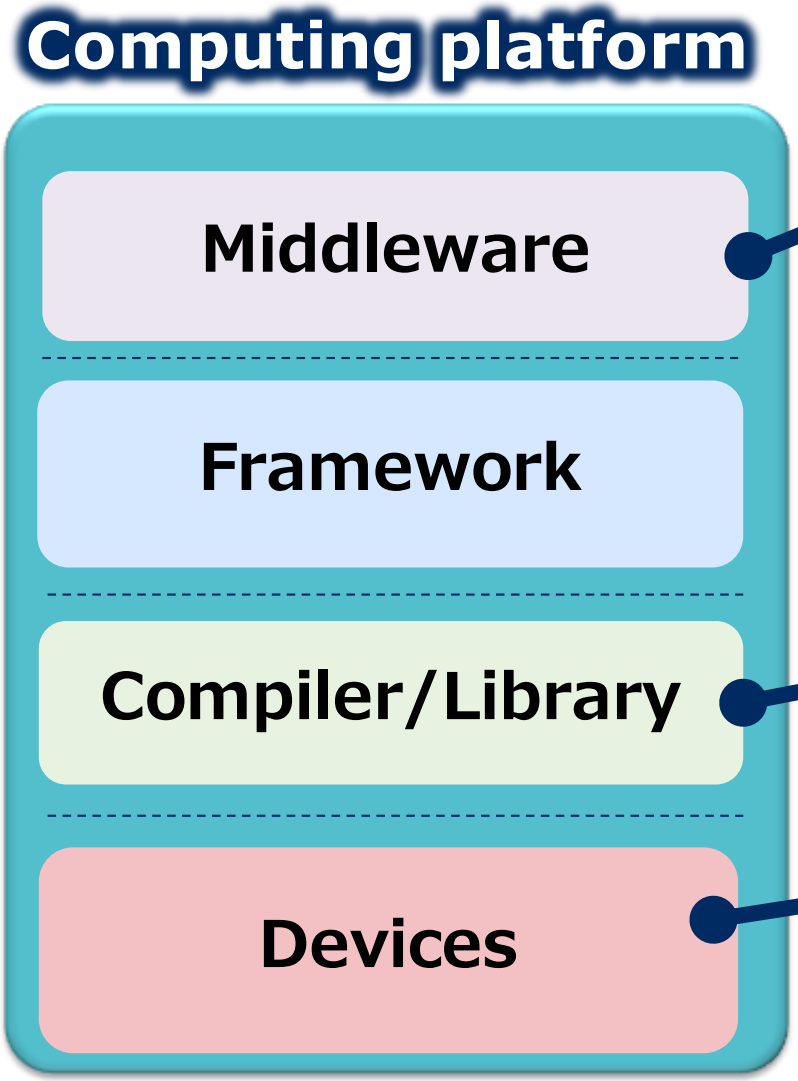
Performance CPU vs FPGA+HLS

C code for Graphics containing branches

```
start;
if (....)
  {filter1;}
if (......)
  {fileter2;}
{Statistic Calc;}
{Decision;}
If (....)
  {complex graphics;}
ending;
```



Our activities for the computing platform for AI

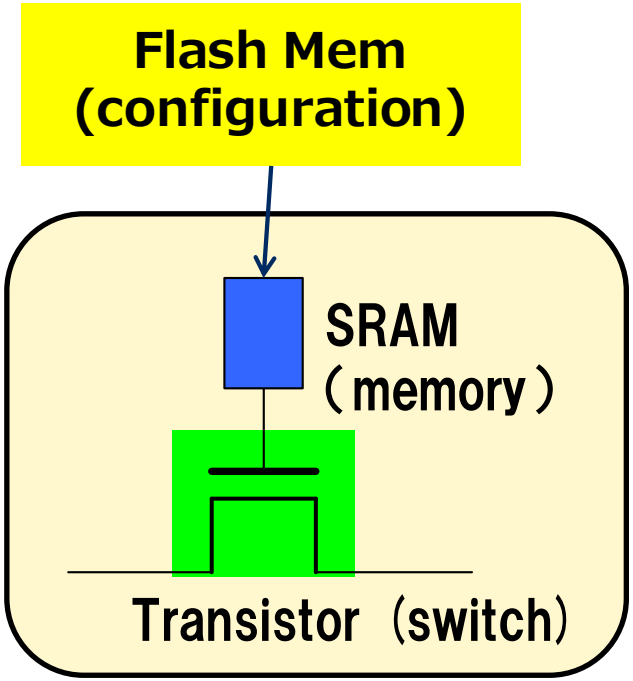


(1) Data management
- Data clustering in D/B

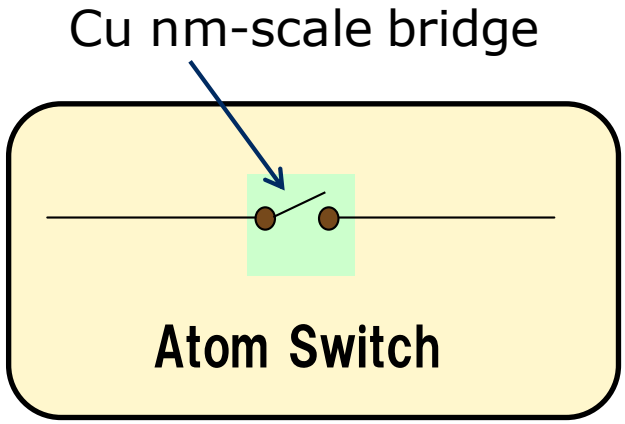
(2) Heterogeneous computing
design tools, libraries

(3) New FPGA
Brain type device

Non volatile FPGA using “nano bridge(atom switch)”



SRAM FPGA

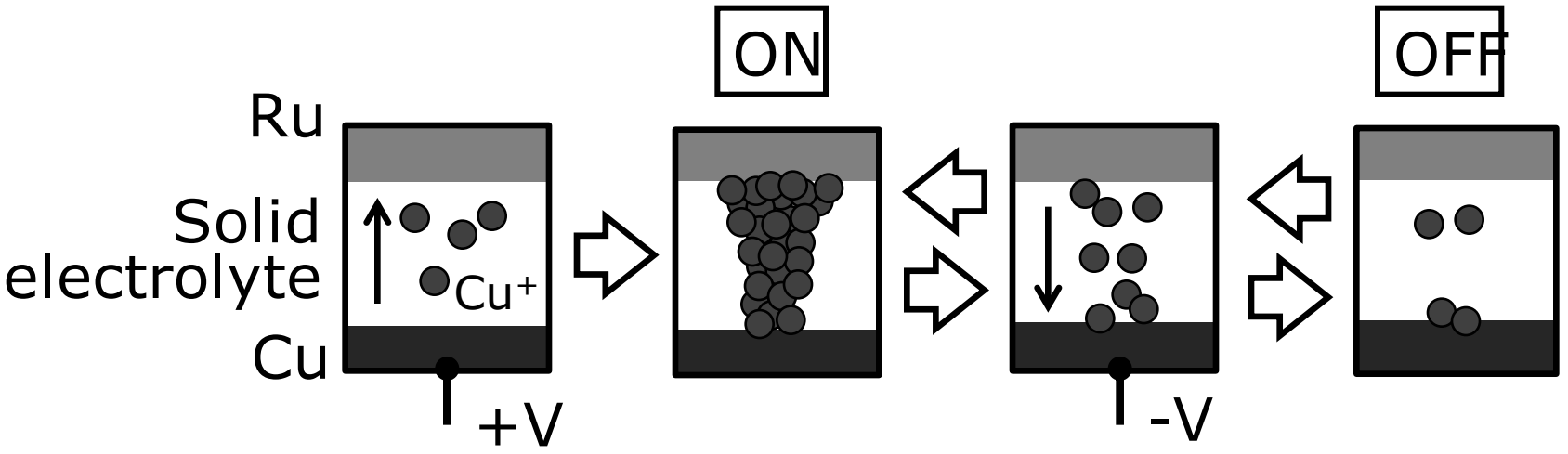


NV-FPGA

Small & Low Power

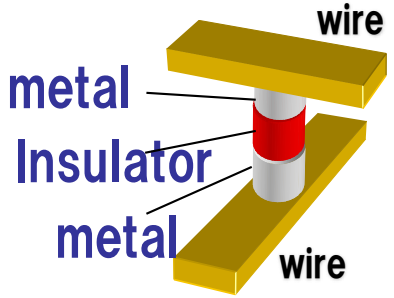
Non Volatile wiring with “Atom Switch(nano bridge)”

- Resistive-switch : Nanometer-scale Cu bridge forms between two electrodes via electrochemical reaction.



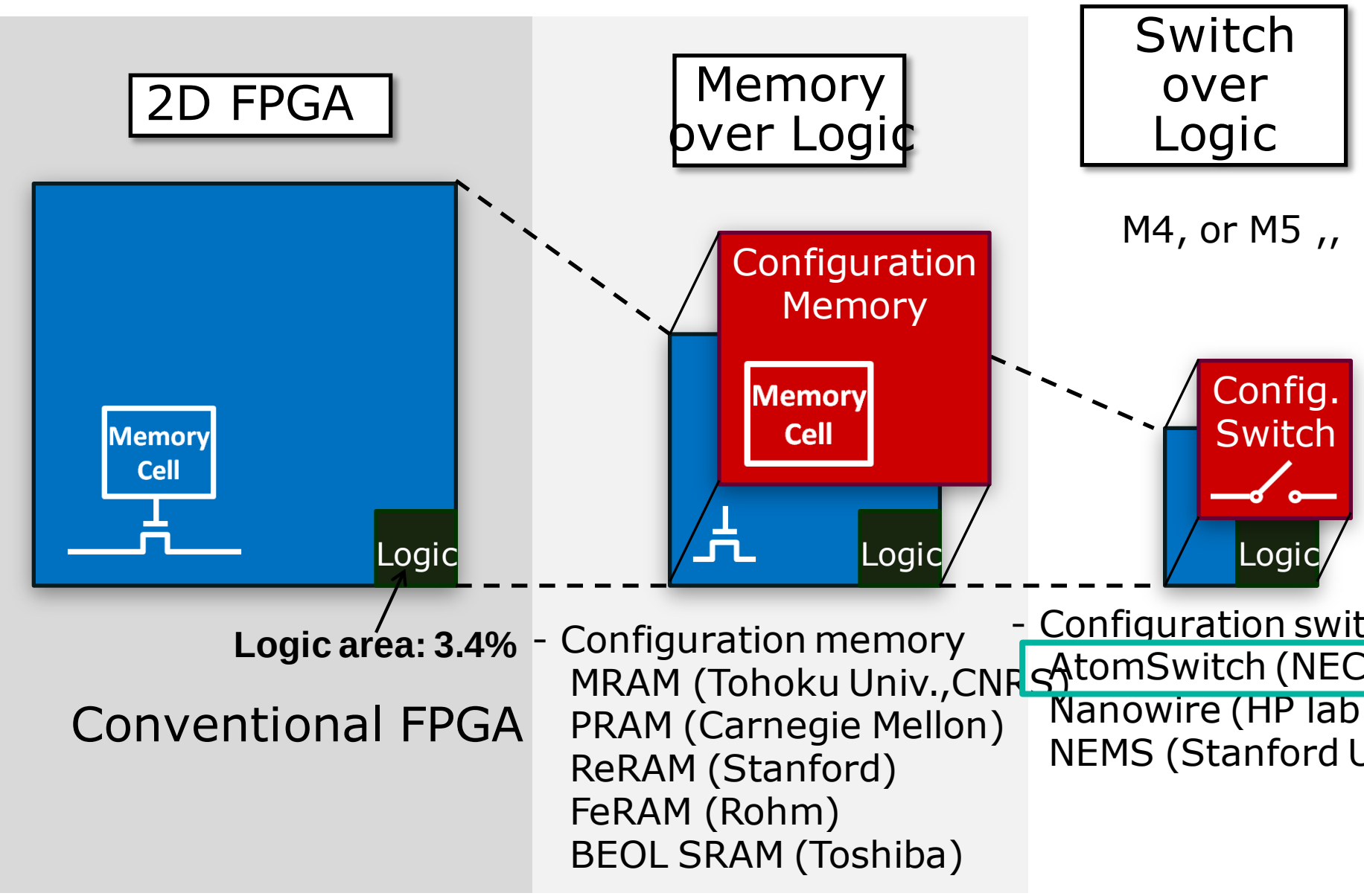
Features

- High ON/OFF conductance ratio (>10^4)
- Nonvolatile
- Rewritable (>10^3)
- Small size



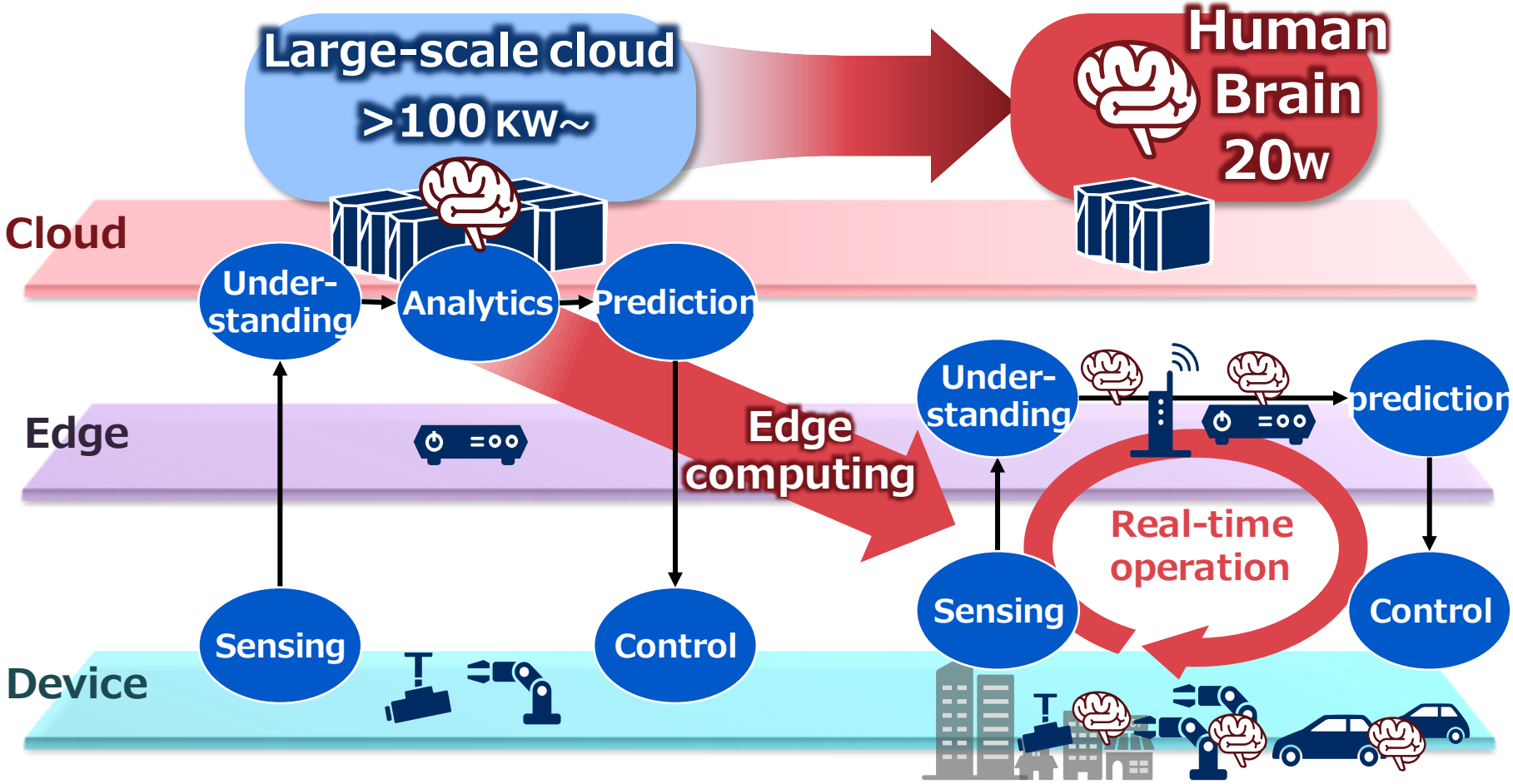
Courtesy of LEAP

“Switch Over Logic” ; do not use transistor for switch



Brain inspired computing (mimic human's brain)

Different type of computing platform for AI is required for resolving various social issues/problems



Open innovation for brain inspired computing

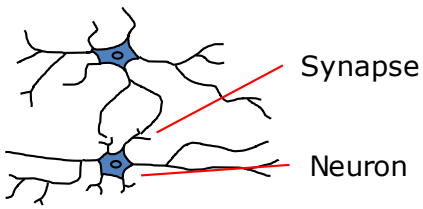
Brain-morphic AI to resolve social issues (Univ. of Tokyo)

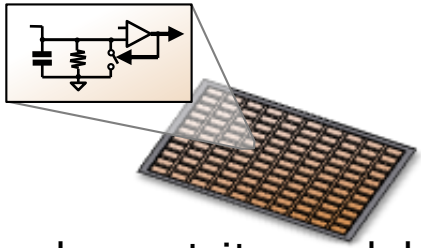
- Investigating a novel computing system by mimicking the behavior of neurons and synapses in the brain.
- Developing the brain-morphic system which can execute fast intellectual-and-autonomous information processing with low energy.

(Already talked in Prof. Kohno's presentation)

https://www.iis.u-tokyo.ac.jp/en/research/department_center/ai/

Modeling brain's behavior at neuron/synapse level





Implement its model as analog circuits

Will deploy it on devices

NEC brain inspired computing Research Alliance Laboratories (Osaka Univ.)

- Investigating novel information processing technology learned from the characteristics of the brain including its environmental adaptability, cognition, and judgment.

http://nbic.ist.osaka-u.ac.jp/

Summary

AI, Iot acceleration

1. FPGA is a good choice for low latency area
2. modify algorithm for FPGA nature
   (e.g. precision reduction, float -> binary)
3. effective usage of **HLS+**FPGA
   free from control dependencies
   (complex control algorithm is good for FPGA+HLLS
   though. **bad for FPGA+RTL**)
   partition arrays into many memories
   (e.g ary[64][128][1024] -> 516 memories)
4. New Devices


Orchestrating a brighter world
NEC

Orchestrating a brighter world

NEC brings together and integrates technology and expertise to create the ICT-enabled society of tomorrow.
We collaborate closely with partners and customers around the world, orchestrating each project to ensure all its parts are fine-tuned to local needs.

Every day, our innovative solutions for society contribute to greater safety, security, efficiency and equality, and enable people to live brighter lives.