Acceleration Technology to Realize Super Resolution Processing of Scanned Image for MFP
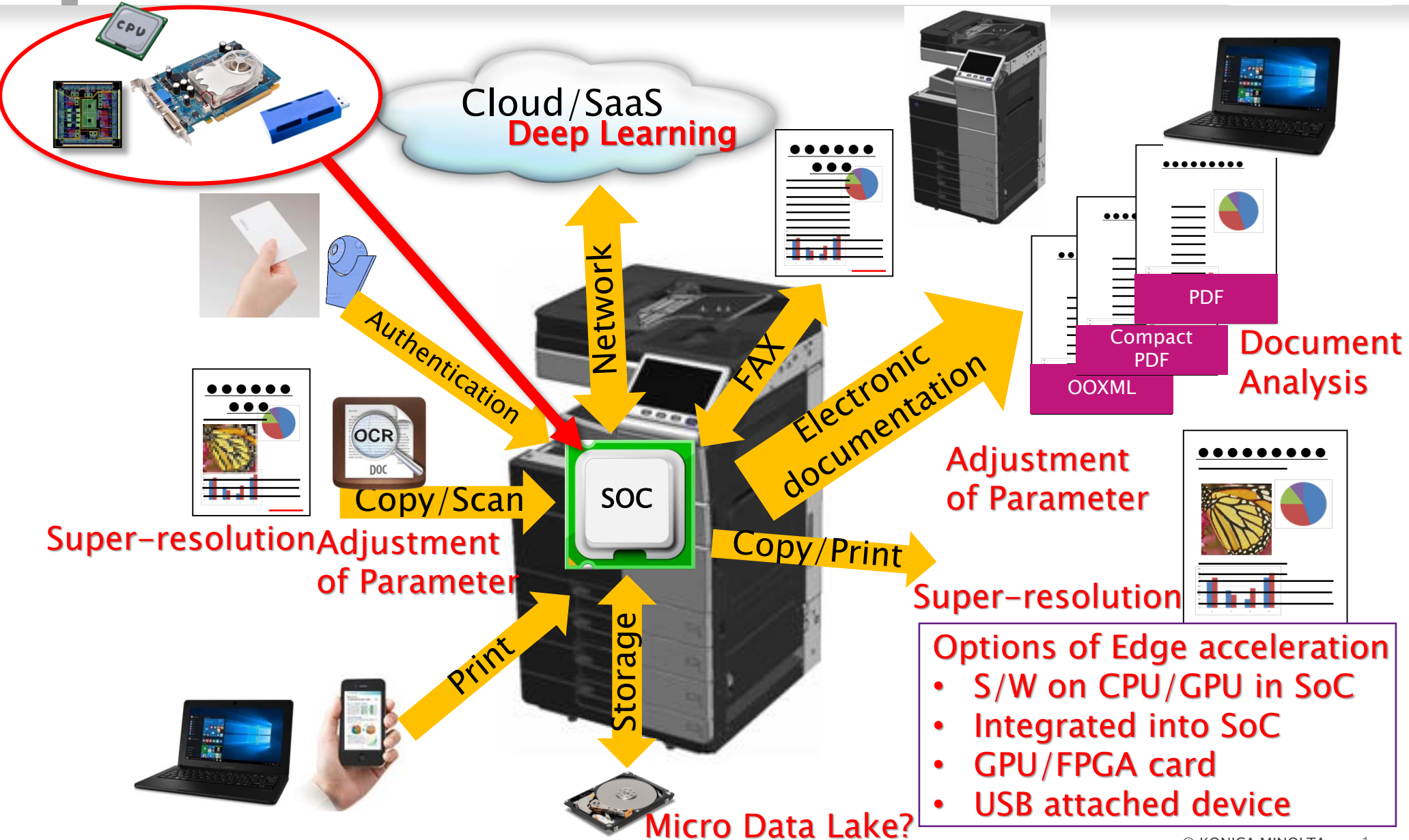
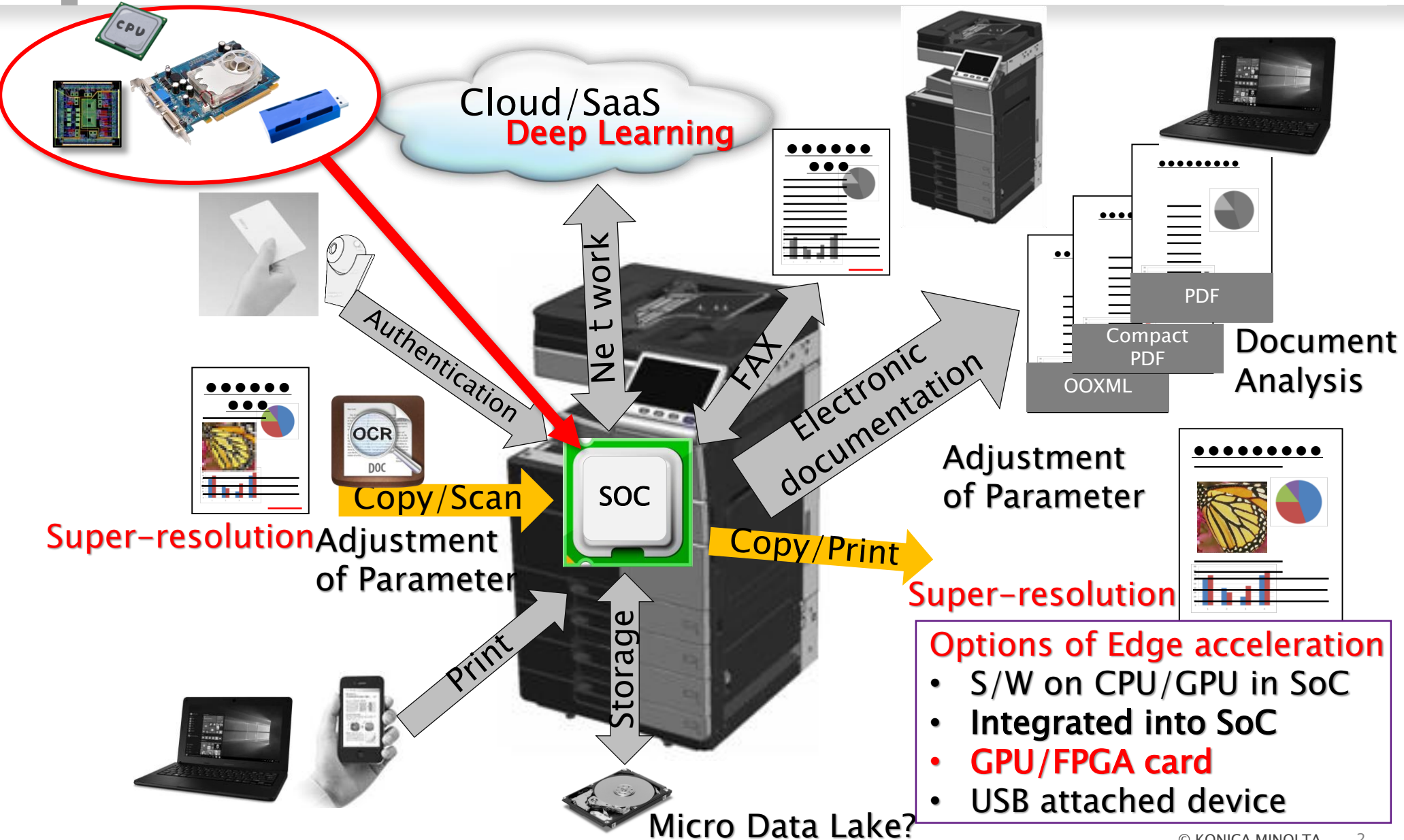Masahiro MURAKAMI

Yuji OKAMOTO

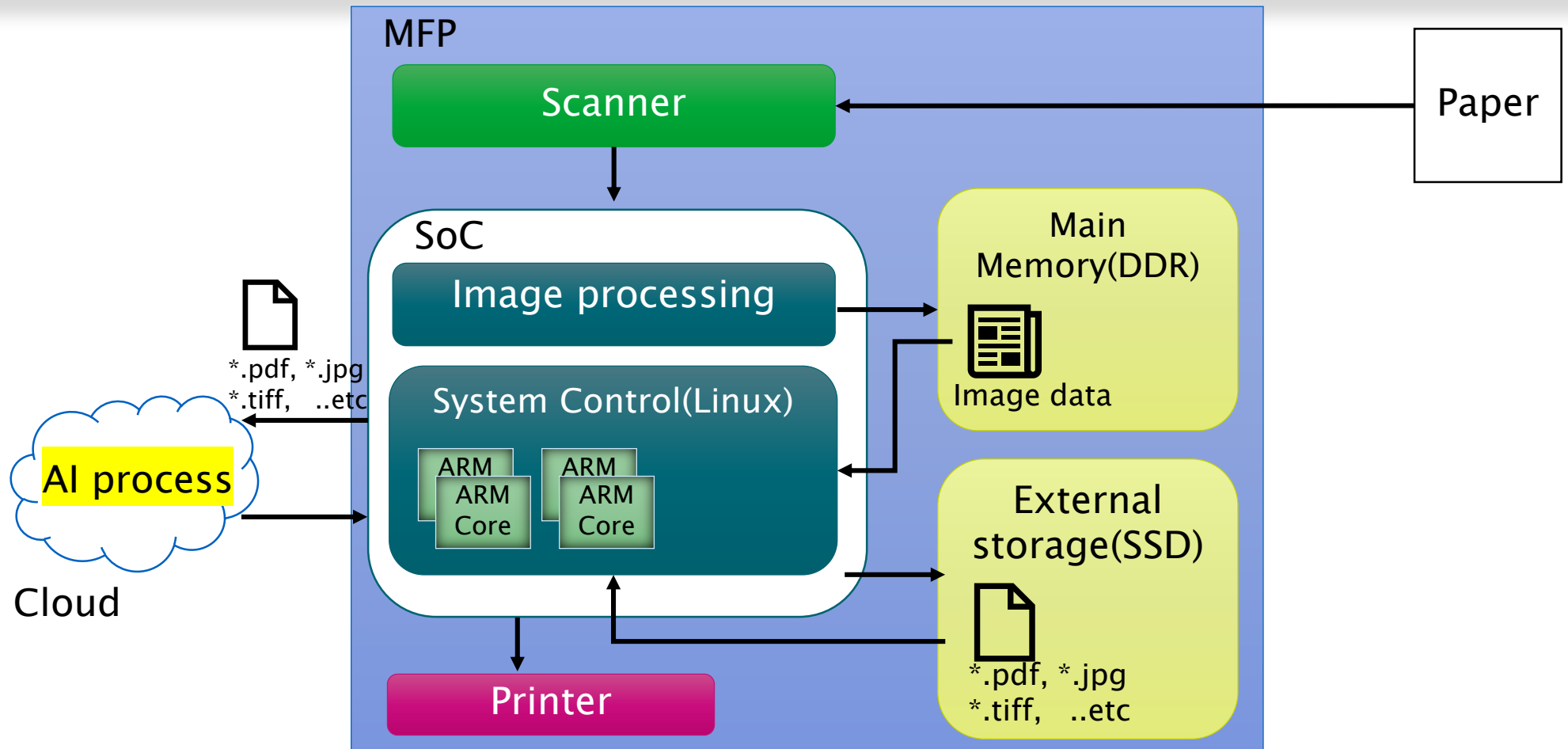KONICA MINOLTA, Inc.

July 8th, 2019

MPSoC'19

Giving Shape to Ideas

Evolution of next-generation MFP by AI

Cloud/SaaS
Deep Learning

Network

FAX

Electronic documentation

Authentication

Copy/Scan

OCR DOC

Super-resolution

Adjustment of Parameter

Print

Storage

SOC

Copy/Print

Document Analysis

PDF
Compact PDF
OOXML

Adjustment of Parameter

Super-resolution

Options of Edge acceleration
- S/W on CPU/GPU in SoC
- Integrated into SoC
- GPU/FPGA card
- USB attached device

Micro Data Lake?

Evolution of next-generation MFP by AI

Cloud/SaaS
Deep Learning

Network

Authentication

FAX

Electronic documentation

Copy/Scan

SOC

Copy/Print

OCR DOC

Super-resolution Adjustment of Parameter

Print

Storage

Micro Data Lake?

PDF
Compact PDF
OOXML

Document Analysis

Adjustment of Parameter

Super-resolution

Options of Edge acceleration
- S/W on CPU/GPU in SoC
- **Integrated into SoC**
- **GPU/FPGA card**
- USB attached device

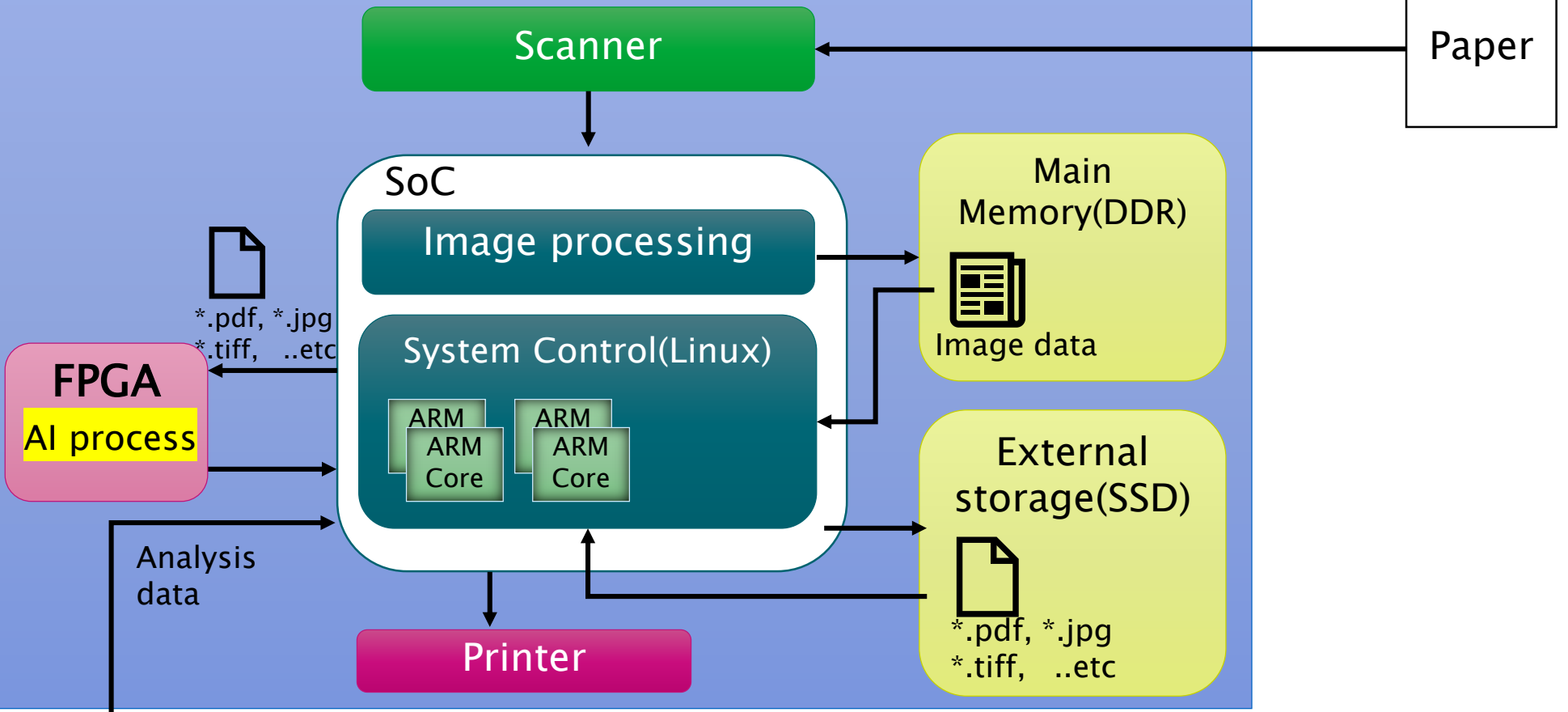# MFP AI system(Current generation)

In current, AI process is implemented in cloud.

However, there are some concerns.

1. privacy concerns
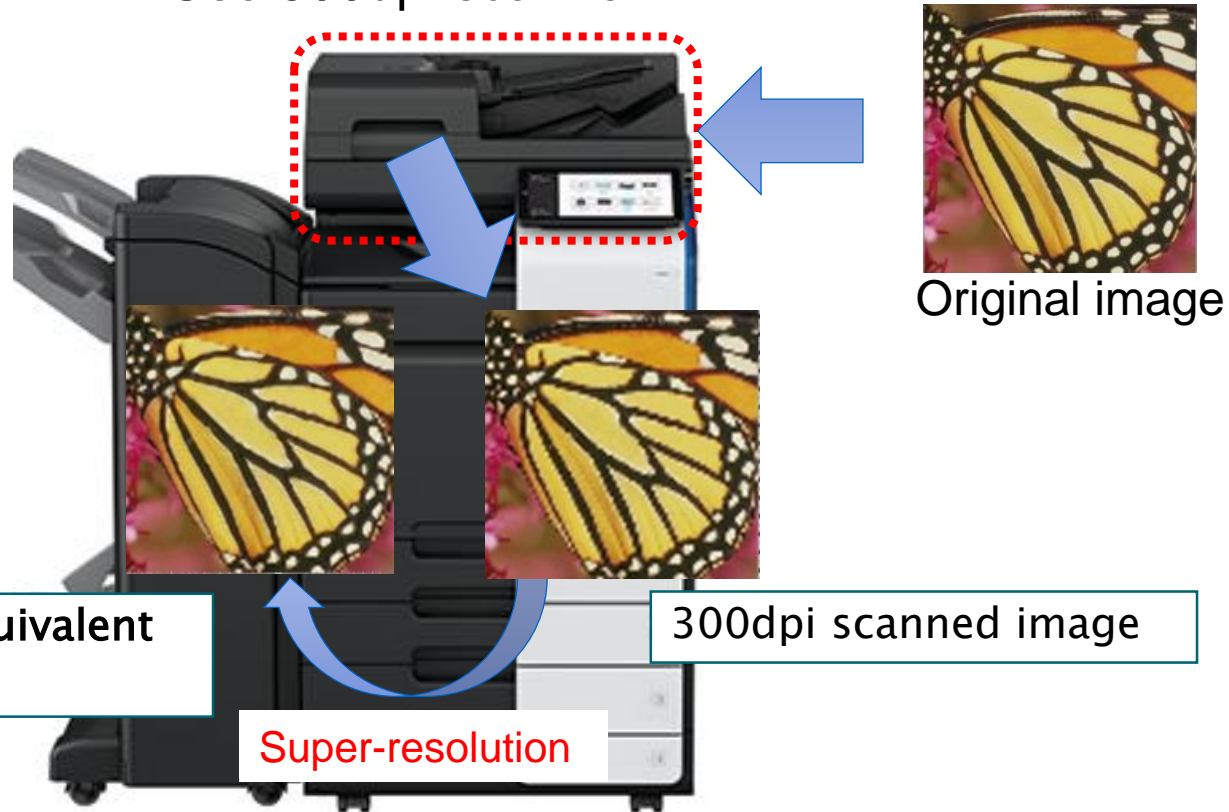2. Realtime/responsiveness concerns
3. Uploading cost

MFP AI system(Current generation)

- As solution for those concern, we are considering AI process on edge.
- The AI accelerator option is GPU, FPGA, ASSP, etc..
- At this time, our first prototype is FPGA at point of power and cost view.

Use case of AI feature for MFP

Use 300dpi scanner

Original image

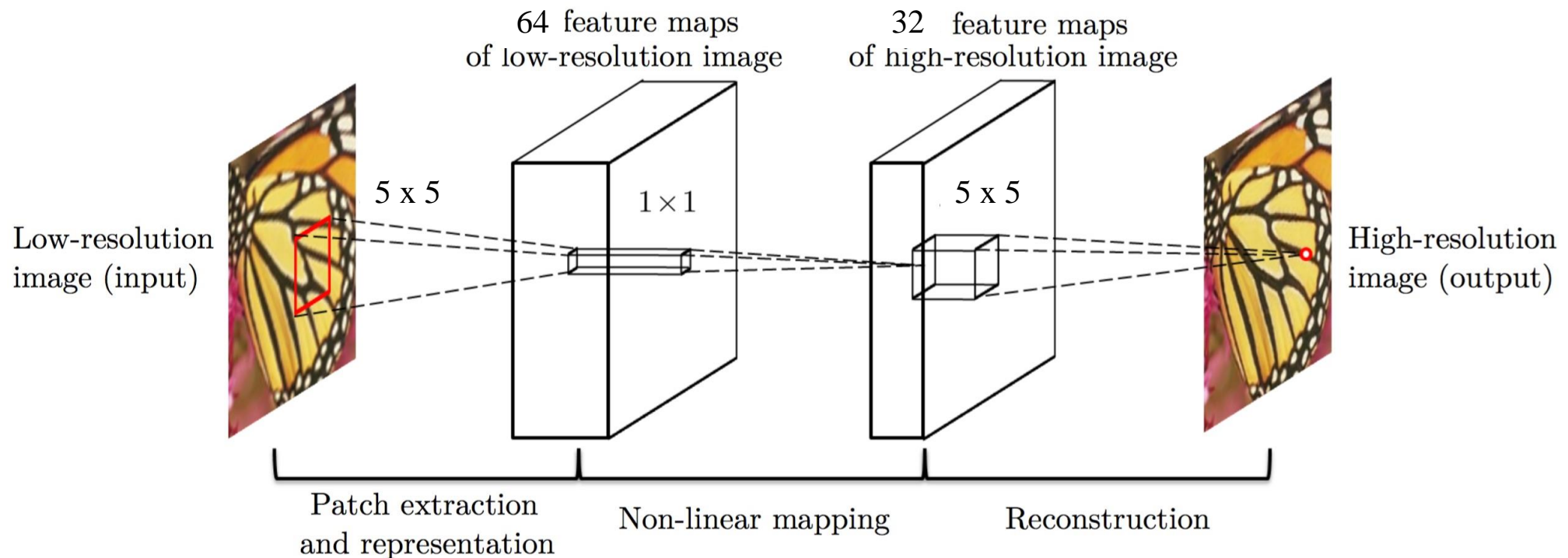Generate 600dpi equivalent scanned image
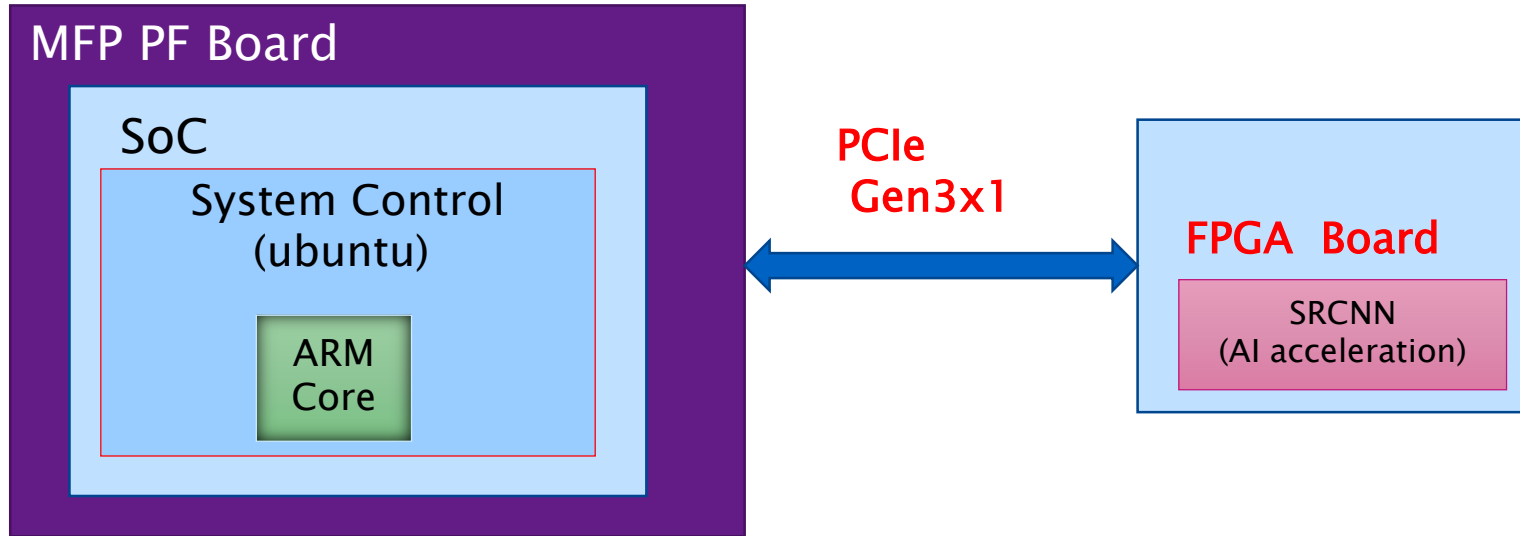
300dpi scanned image

Super-resolution

- Cost Down
  - Using cheap scanner, but enable to get high quality image.
- Function quality Up
  - Printing quality up for generation copy.
  - Saving external storage(small file size output).

Target Neural Network is SRCNN.
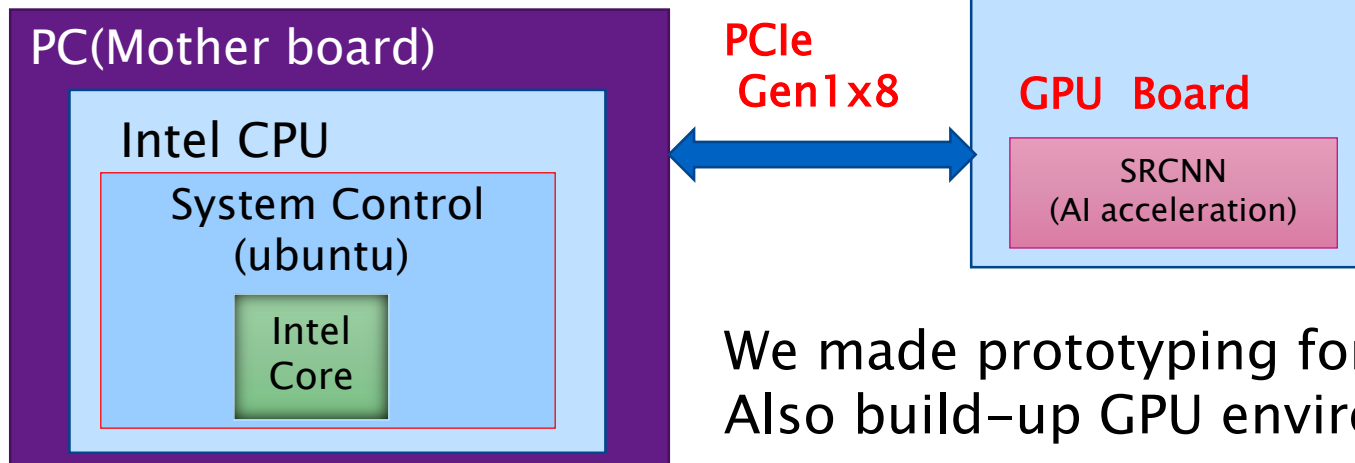
- – SRCNN was published in 2014 ECCV.
- – we can get a better quality of a larger image from a small image originally.

64 feature maps of low-resolution image

32 feature maps of high-resolution image

Low-resolution image (input)

5 x 5   1×1   5 x 5

High-resolution image (output)

Patch extraction and representation   Non-linear mapping   Reconstruction

Trial Evaluation Environment

■MFP PF(SoC) + FPGA(PCIe Gen3 x1)

MFP PF Board

SoC

System Control (ubuntu)

ARM Core

PCIe Gen3x1

FPGA Board

SRCNN (AI acceleration)

■PC+GPU(PCIe Gen1 x8)

PC(Mother board)

Intel CPU

System Control (ubuntu)

Intel Core
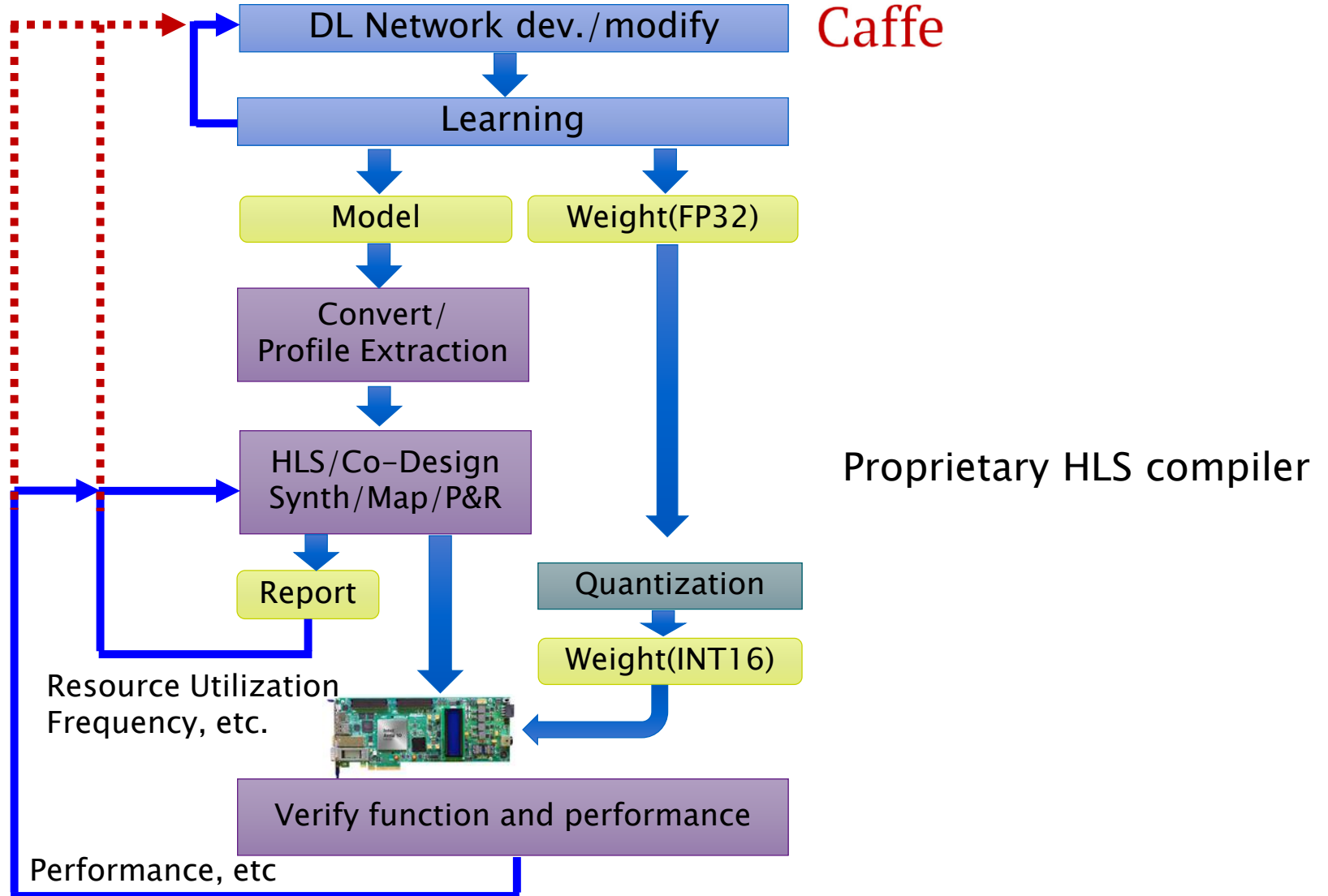
PCIe Gen1x8

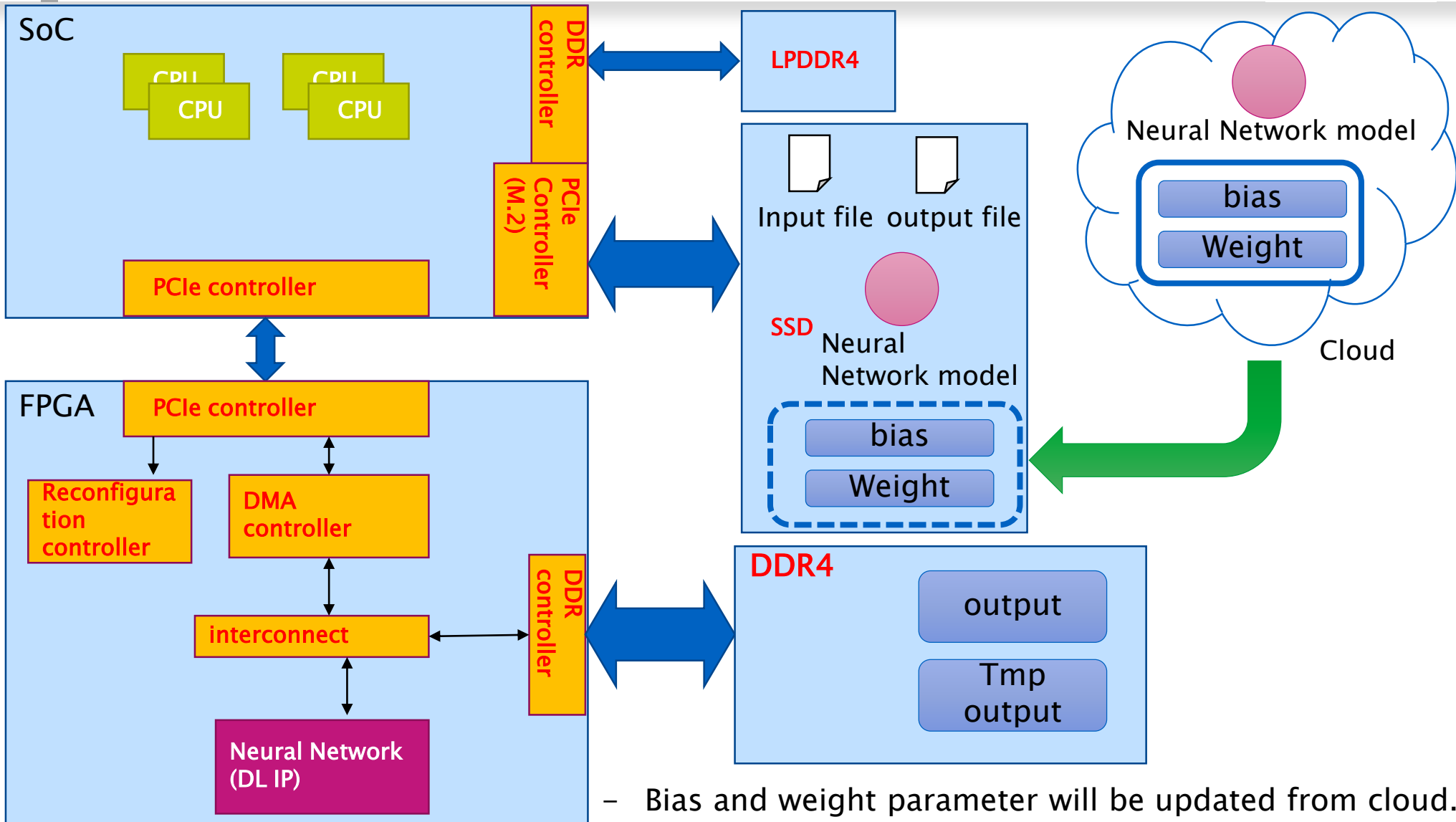**GPU Board**

SRCNN (AI acceleration)

We made prototyping for FPGA AI acceleration. Also build-up GPU environment for comparing.

Development Flow

In our system, we can get FPGA AI accelerator from DL network model file.

Caffe

Proprietary HLS compiler

- DL Network dev./modify
- Learning
  - Model
    - Convert/Profile Extraction
      - HLS/Co-Design Synth/Map/P&R
        - Report
  - Weight(FP32)
    - Quantization
      - Weight(INT16)
- Verify function and performance

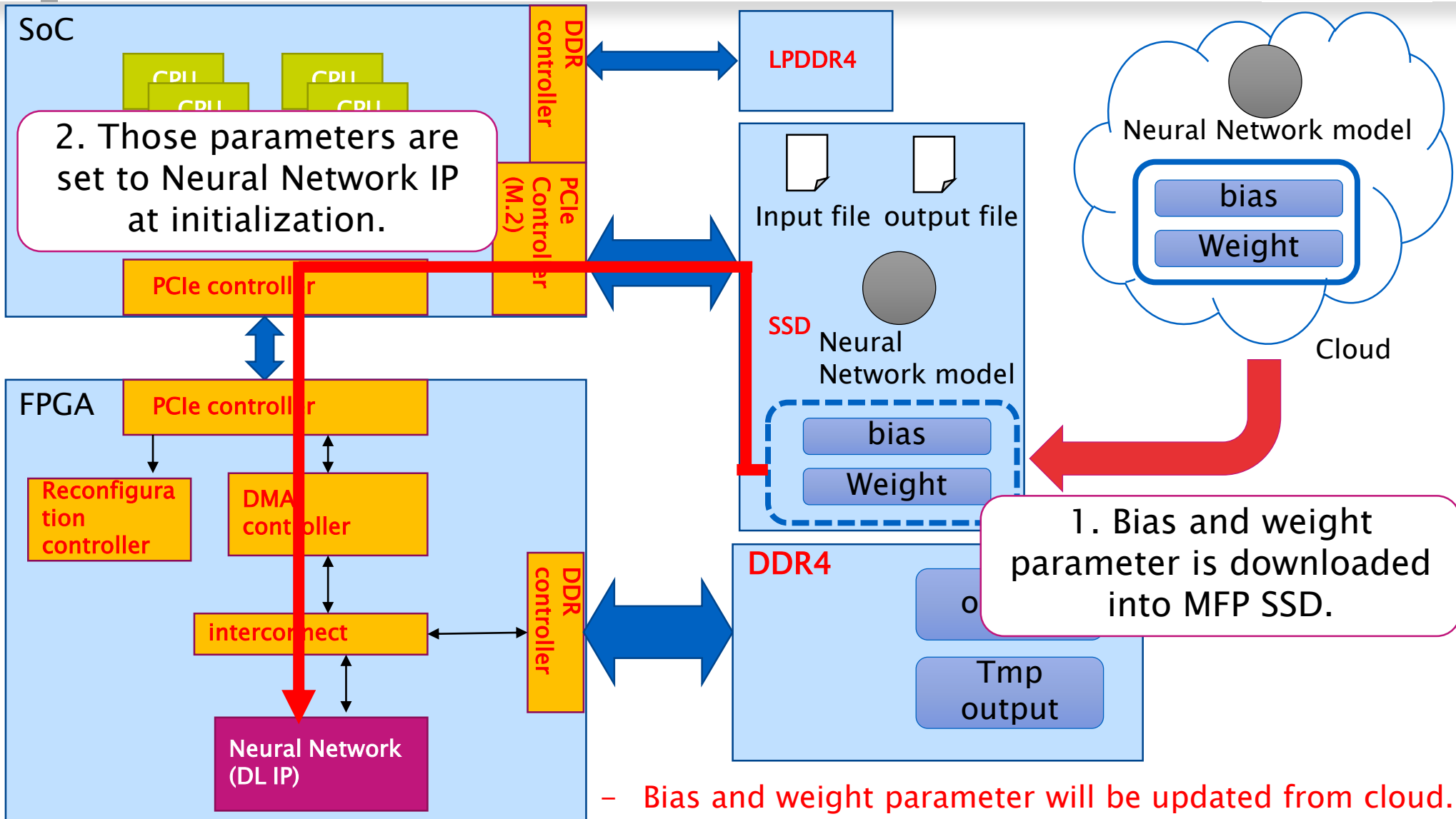Resource Utilization Frequency, etc.

Performance, etc

Our Target System Structure

- Bias and weight parameter will be updated from cloud.
- Neural Network design on cloud will be switched by reconfiguration controller.

Our Target System Structure

SoC

CPU CPU CPU CPU

DDR controller

LPDDR4

2. Those parameters are set to Neural Network IP at initialization.

PCIe Controller (M.2)

PCIe controller

Input file   output file

SSD

Neural Network model

bias

Weight

Cloud

Neural Network model

bias

Weight

FPGA

PCIe controller

Reconfiguration controller

DMA controller

interconnect

DDR controller

DDR4

o

Tmp output

1. Bias and weight parameter is downloaded into MFP SSD.
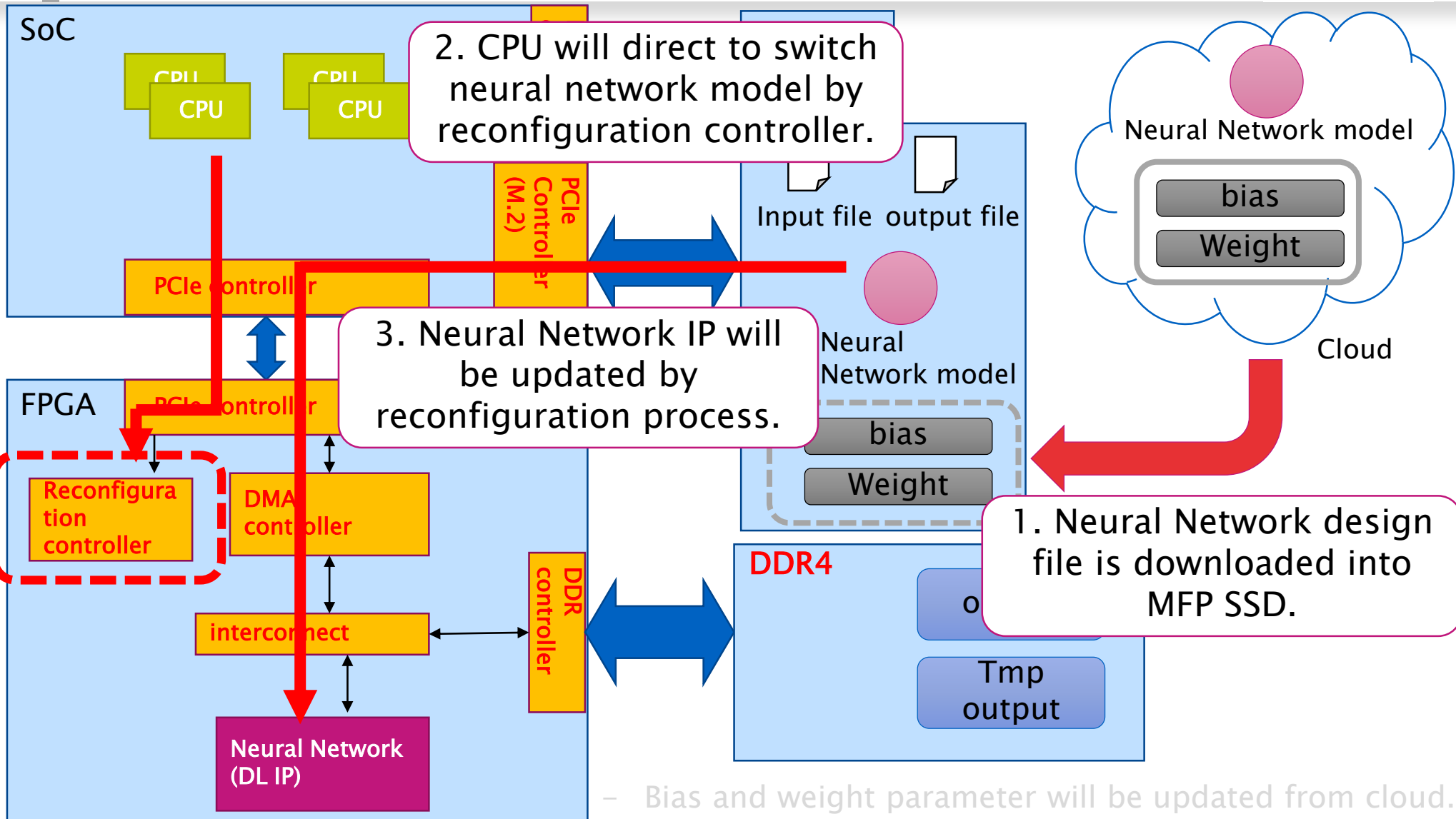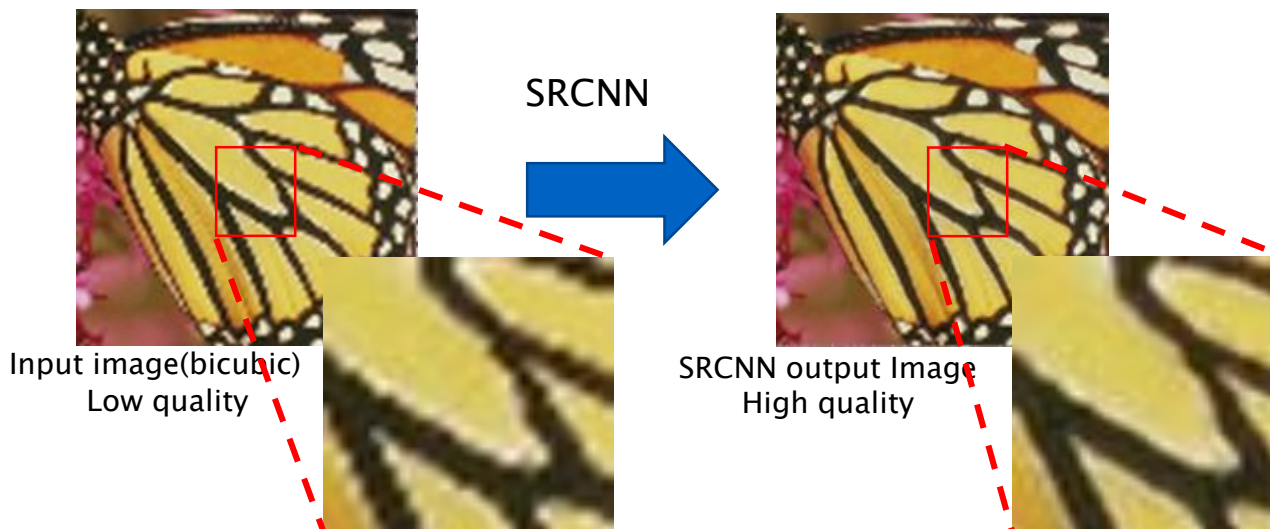
Neural Network (DL IP)

– Bias and weight parameter will be updated from cloud.
– Neural Network design on cloud will be switched by reconfiguration controller.

Our Target System Structure

SoC

CPU CPU

CPU CPU

PCIe controller

PCIe Controller (M.2)

Input file output file

2. CPU will direct to switch neural network model by reconfiguration controller.

Neural Network model

bias

Weight

Cloud

FPGA

PCIe controller

3. Neural Network IP will be updated by reconfiguration process.

Neural Network model

bias

Weight

Reconfiguration controller

DMA controller

DDR controller

interconnect

Neural Network (DL IP)

DDR4

o

Tmp output

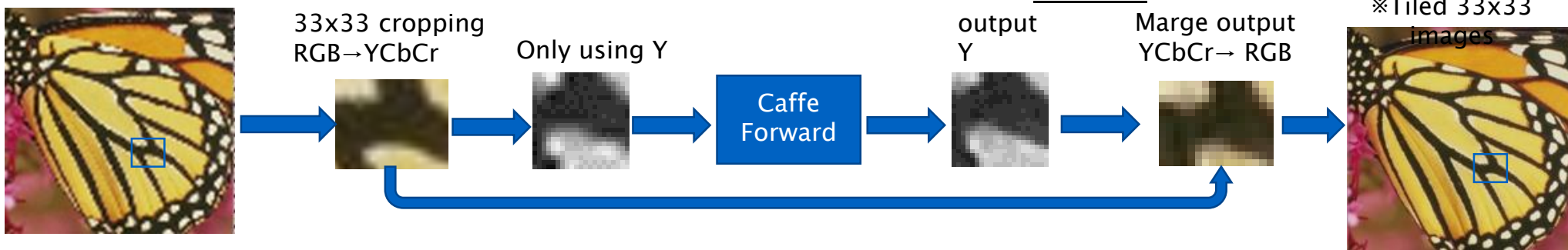**1. Neural Network design file is downloaded into MFP SSD.**

– Bias and weight parameter will be updated from cloud.
– Neural Network design on cloud will be switched by reconfiguration controller.

Inference Result

SRCNN

Input image(bicubic)
Low quality

SRCNN output Image
High quality

It seems getting good result.

[Reference]: Inference flow

33x33 cropping
RGB→YCbCr

Only using Y

Caffe
Forward

output
Y

Marge output
YCbCr→ RGB

※Tiled 33x33 images

Output Image Quality

◆ We evaluated image quality by PSNR and SSIM.

Input image
(Bicubic)

Output image
(CPU_FP32)
No quantization

Output image
(FPGA_INT16)
After quantization
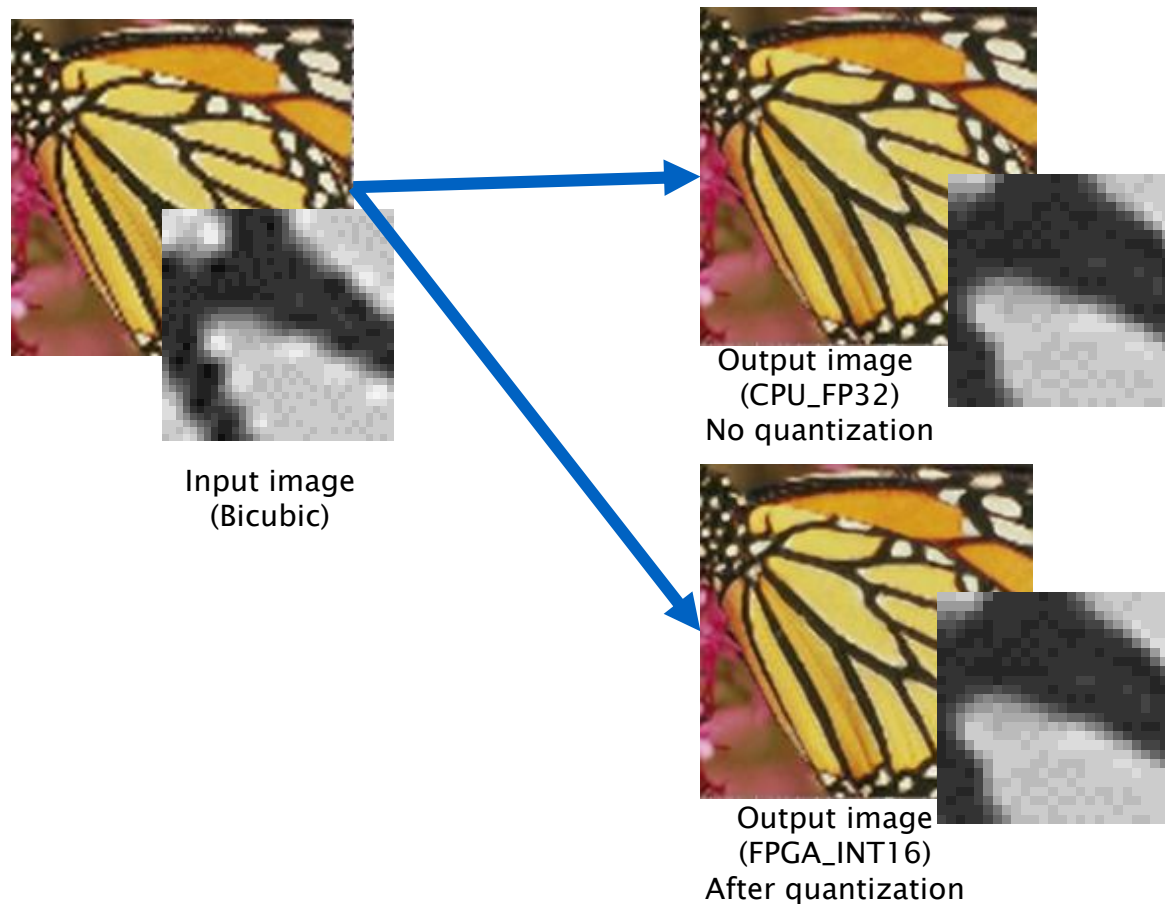
Image data	PSNR[dB]
Input image Bicubic	23.4273
Inference by CPU (FP32)	24.2970
Inference by FPGA(INT16)	24.2946

Image data	SSIM
Input image Bicubic	0.784
Inference by CPU (FP32)	0.810
Inference by FPGA(INT16)	0.810

- The quantized INT16 parameter is very little influence for output image data.
- Our Proprietary HLS compiler works well.

◆ Performance is insufficient with only CPU. AI accelerator is required for super resolution of scanner input size.

◆ FPGA accelerates super resolution process, but A4 size super resolution process takes more than a minute, and it is not realistic to use as it is.
We are required continuous effort for performance improvement for AI accelerator.
  - Neural Network optimization. (changing block size larger)
  - Approach of using quantization to 8bit.
  - Cropping only low quality region in A4 scan image and only input a part of images.

Inference process[s] (image size A4(7016x4961)

Future plan

CPU@3.5GHz
Not optimized   FPGA@166MHz   GPU Max 13.4TFLOP

Movie

Conclusion and future plan
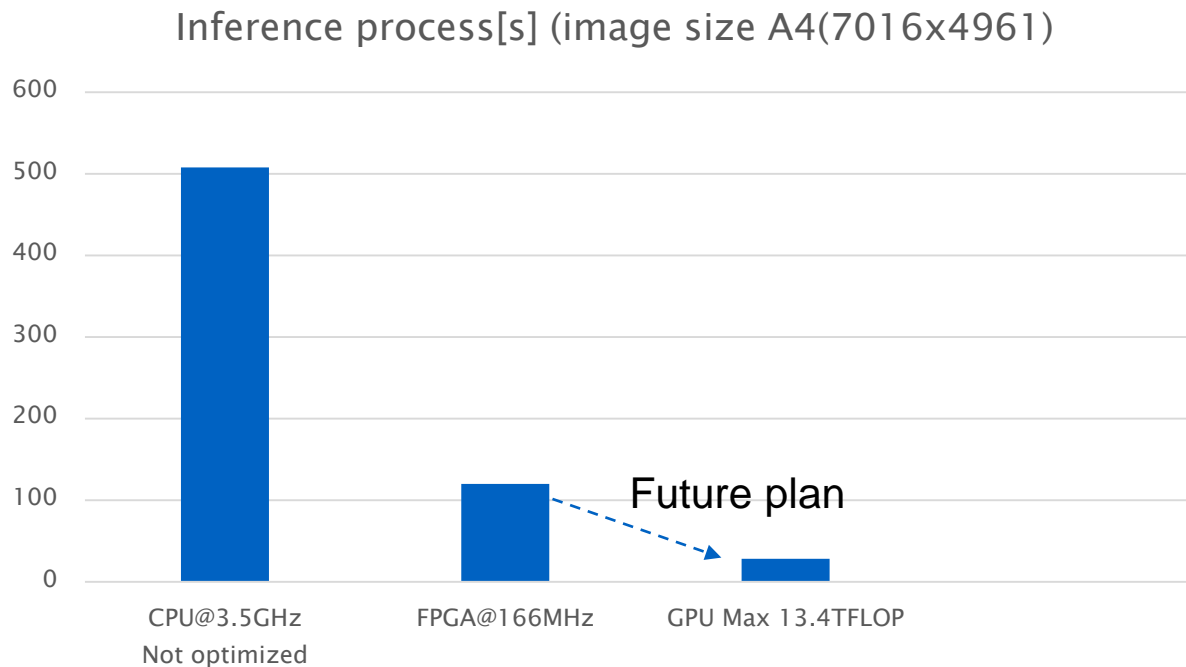
◆ We got FPGA AI accelerator from DL network model file by proprietary HLS compiler.

◆ In our system, reconfiguration controller can switch neural network design in FPGA.

◆ Quantization from FP32 to INT16 does not affect to image quality.

◆ Performance is insufficient with only CPU. AI accelerator is required for super resolution of scanner input size.

◆ We are considering multiple AI accelerator option, GPU, FPGA and ASSP/IP.

◆ FPGA and GPU accelerates super resolution process but, it was slower than we expected. We are required continuous effort for performance improvement for FPGA AI accelerator.
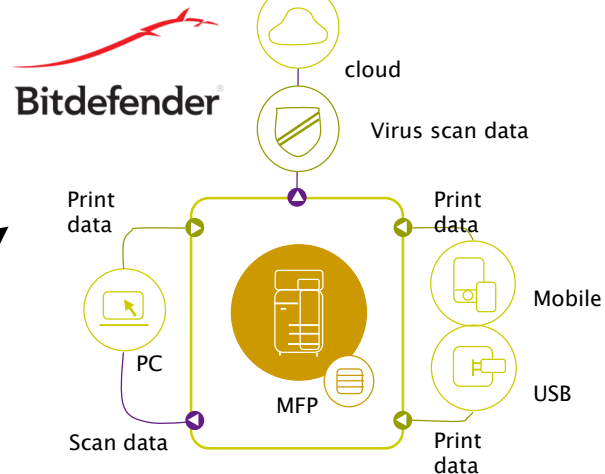
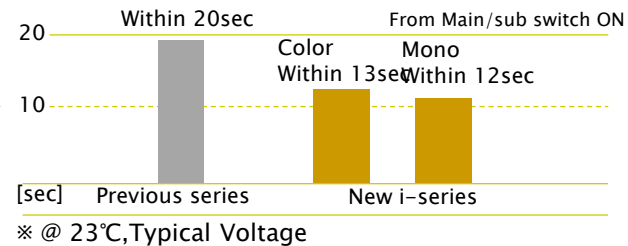# *Thank you!*

KONICA MINOLTA's New MFP i-SERIES

■Virus scan

Bitdefender

cloud

Virus scan data

Print data

Print data

Mobile

PC

USB

MFP

Scan data

Print data

CA72 x4

Cortex A72 Quad Core

High performance of the CPU improves user experiences

■Warmup time(previous series VS New i-series)

Within 20sec

From Main/sub switch ON

20

Color Within 13sec

Mono Within 12sec

10

[sec]   Previous series   New i-series

※ @ 23℃,Typical Voltage

■Panel wake-up time
  (previous series VS New i-series)

5

About 4.8sec

About 1.2sec

[sec]   Previous series   New i-series