FUJITSU

shaping tomorrow with you

ISC 2018

DLU™ : Deep Learning Unit

Fujitsu's HPC Development Timeline

FUJITSU

Deep Learning Unit (DLU™)

DLU is a processor designed for deep learning that has the ability to handle large-scale neural networks.

K computer

The K computer is still competitive in various fields; from advanced research to manufacturing.

© RIKEN

SPARC64

SPARC64

SPARC64

ARM

AWARDS Gordon Bell Prize

HPCG

GRAPH 500

Gordon Bell Prize Finalist (2016)

HPCG No.1 (2016)

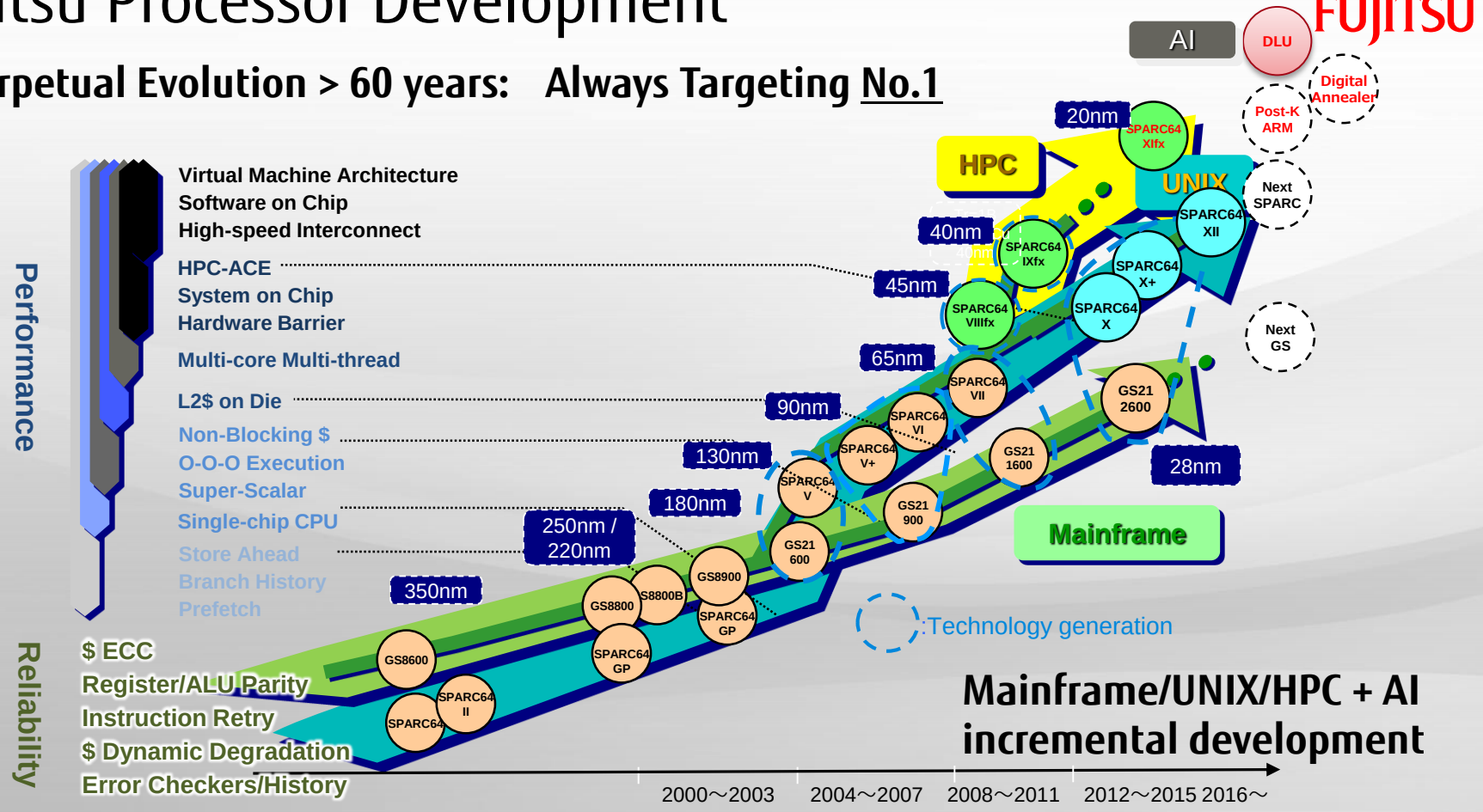
Graph500 No.1 (2016)

Post-K Computer

The post-K is under development to achieve superior application performance.

Fujitsu Processor Development

Perpetual Evolution > 60 years: Always Targeting No.1



DLU: Processor Designed for Deep Learning

FUJITSU

DLU
Deep Learning Unit

Utilizing technologies derived from the K computer

K computer

Features

- Architecture designed for deep learning
- Low-power consumption design

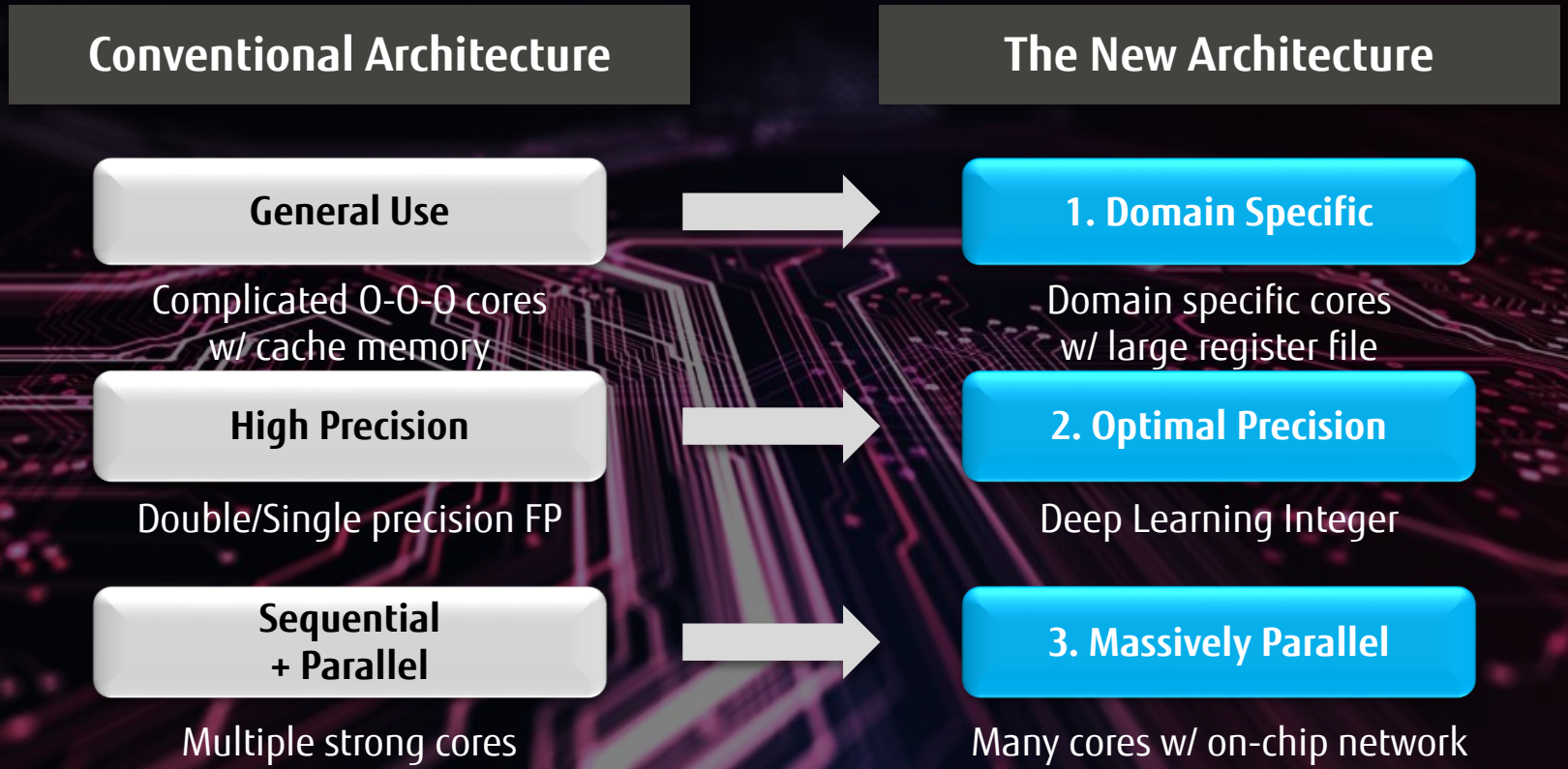
» **Goal: 10x Performance / Watt compared to competitors**

- Scalable design with Tofu interconnect technology

» **Ability to handle large-scale neural networks**

What's the New Architecture for the DLU?

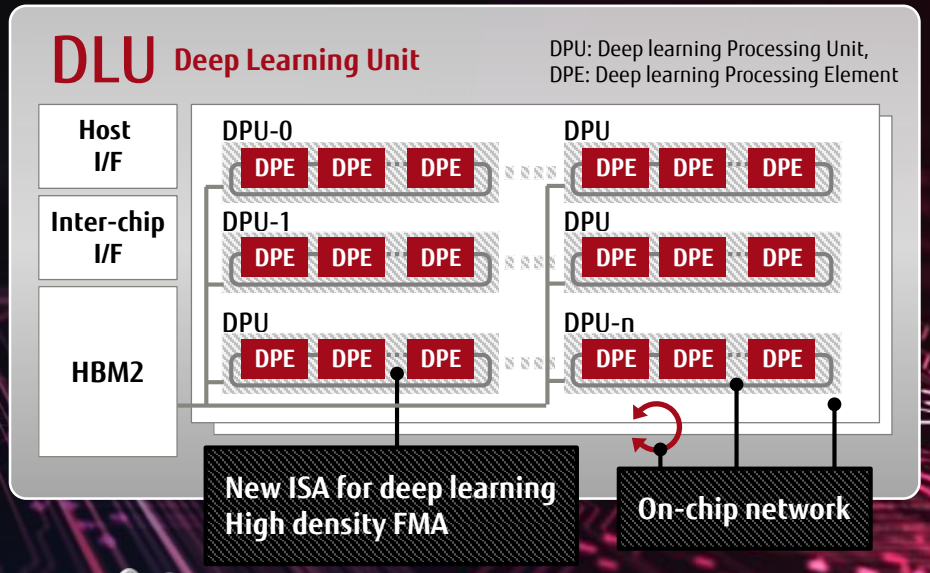
Domain specific, Optimal precision, and Massively parallel.

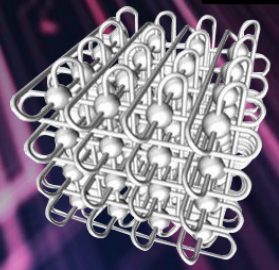


DLU Architecture

FUJITSU

- ISA: Newly developed for deep learning
- Micro-Architecture
  - Simple pipeline to remove HW complexity
  - On-chip network to share data between DPUs
- Utilizes Fujitsu's HPC experience, such as high density FMAs and high speed interconnect
- Maximizes performance / watt





Fujitsu's interconnect technology

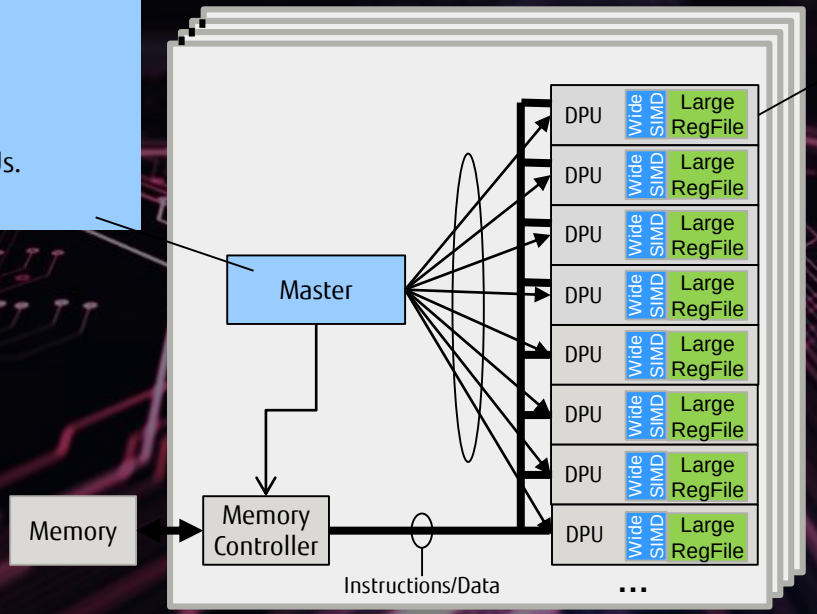
Large scale DLU interconnect through off-chip network

Heterogeneous Cores and Large Register File

The combination of few large core (Master) and many small execution cores (DPU) results in more performance with less power consumption, compared to a conventional homogeneous structure

Master Core:
Memory Access and
DPU control

- Push & Pull
  instructions and data for DLUs.
- Start/stop execution of DLUs

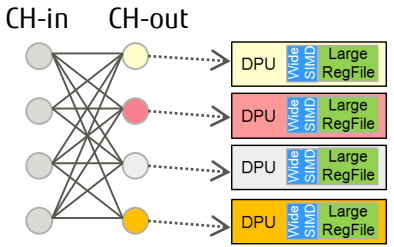


DPU: Execution

- Wide SIMD execution and Large Register File.
- Execute DL operations based on master core's control

How to utilize many DPUs

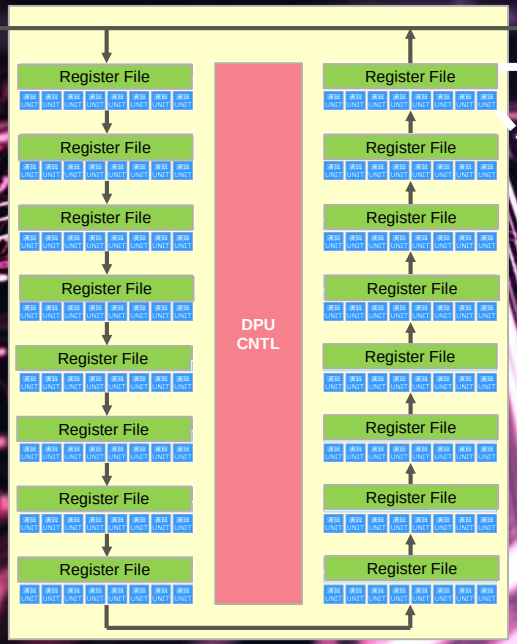
- e.g. convolution layer
  one CH-out / DPU
  multiple batch / DPU



DPE & Large RF (Register File)

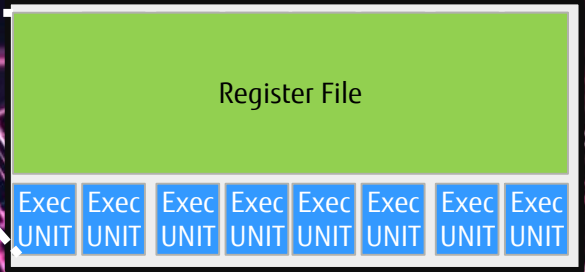
- DPU consists of 16 DPEs connected with on-chip network
- DPE incudes large RF and wide SIMD execution units to realize an efficient Deep Learning engine.
  - RF is fully SW controllable unlike cache to extract full HW potential

DPU: 128 SIMD* / 16DPE



DPE: 8SIMD* with large RF (~100x of typical CPU core)

* For FP32

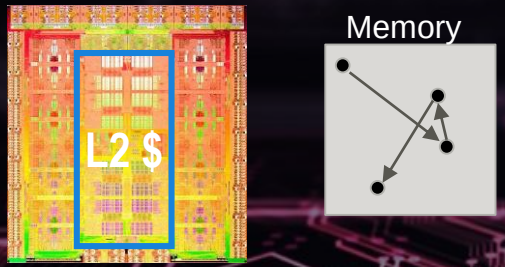


	Name	RF/$ structure
UNIX	SPARC64 XII	RF + $
HPC	SPARC64 XIfx	RF + sector $
AI	DLU	Large RF

More SW controllability

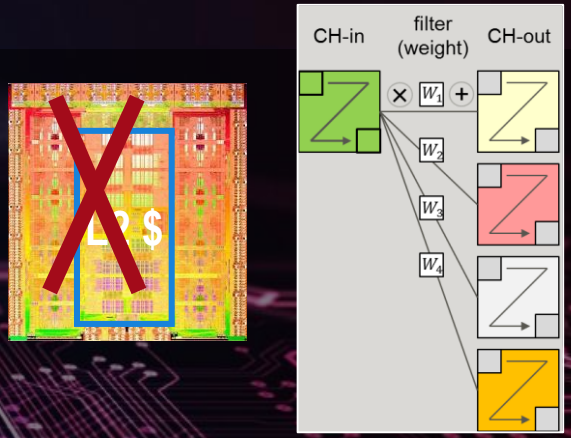
Domain specific architecture - Why cache memory removed -

General Processors



- Complex hardware to achieve high performance for any applications with various data access patterns
- E.g.
  - Large cache memory with cache tags and LRU replacement Unit
  - Hardware Prefetch Engine

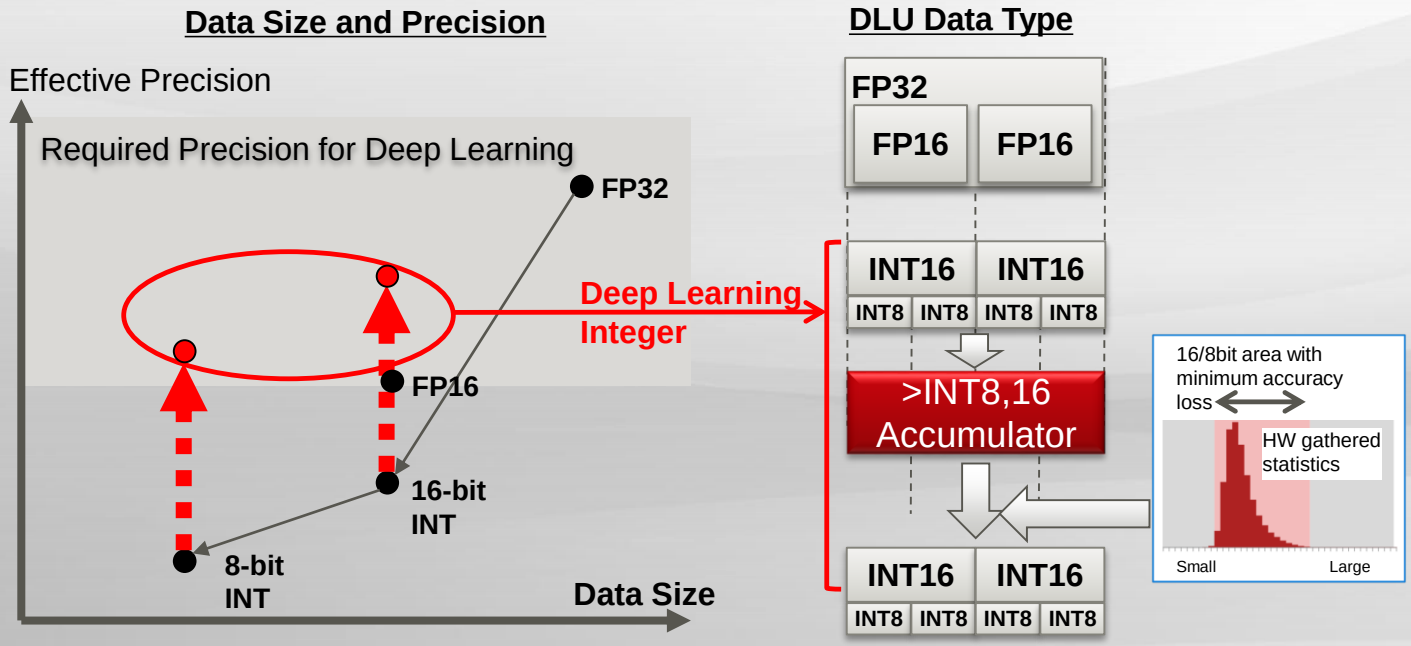
DLU (Domain Specific)



- Simple hardware focusing on simple memory access patterns
- E.g. Convolution Layer
  - CH-in data can be shared among CH-out calculation at all DPUs
  - Memory access patterns are continuously and predictable (software controllable)

DLINT : Deep Learning Integer

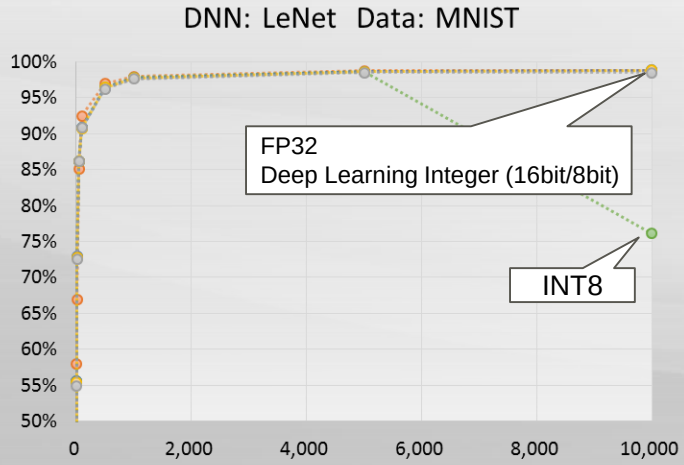
- Fujitsu's "DLINT" realizes necessary accuracy for Deep Learning with only a 16 or 8 bits data size (i.e. less power consumption compared with FP32)
- Training results with DLINT8/16 can be converted to the conventional 8/16-bit INT for inference.

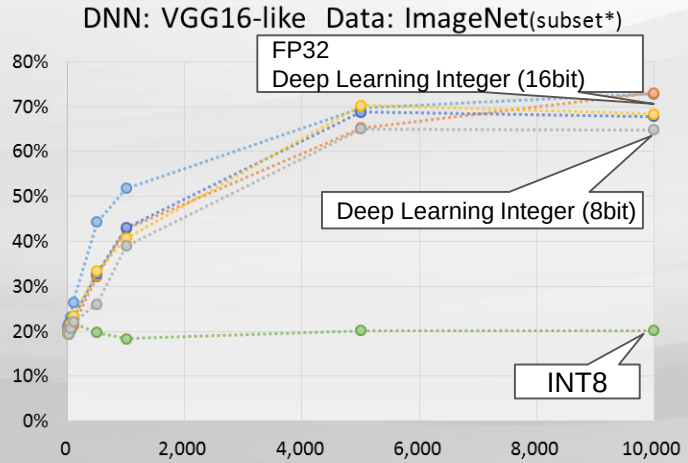


Accuracy of Deep Learning Integer

FUJITSU

DLINT has shown similar accuracy with FP32 precision



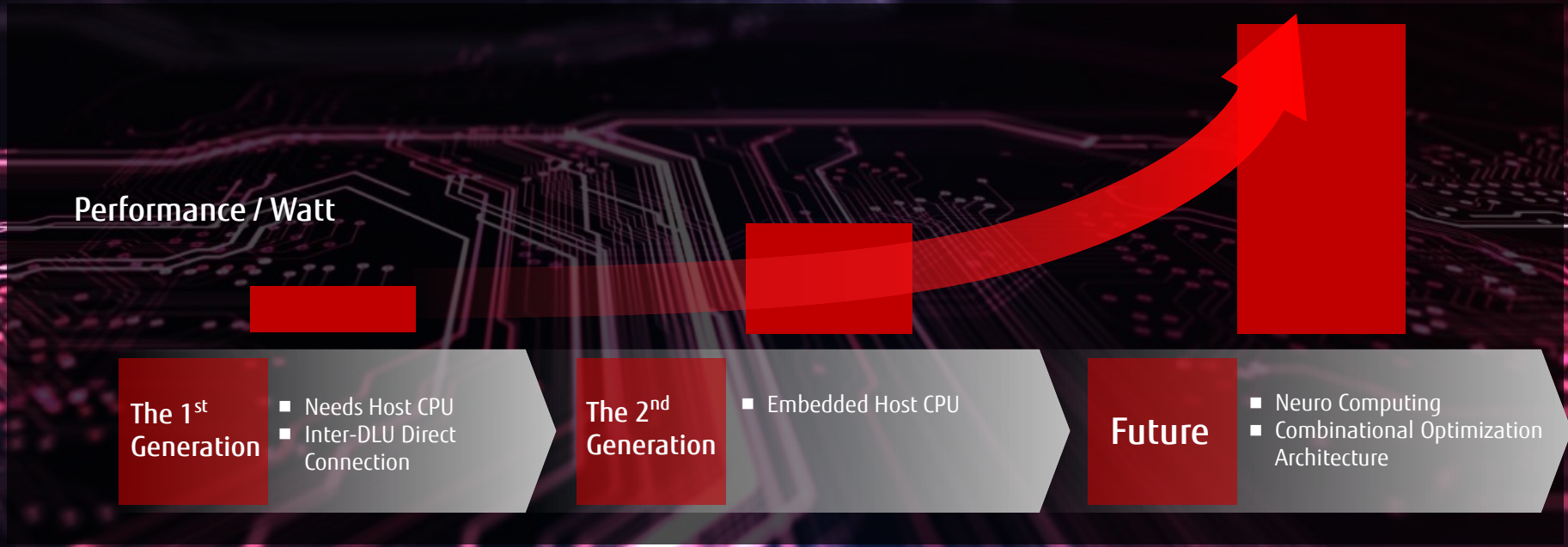


(*) ImageNet(subset): image size=96x96, #categories=25

DLU Roadmap

FUJITSU

- Multiple generations of DLUs over time, as we currently do for HPC/UNIX/Mainframe processors



* Subject to change without notice

AI will be accelerated by three technologies

Quantum Computing

Digital Annealer will identify combinational optimal solutions

Three world-class advanced technologies together will contribute to expansion of customer business

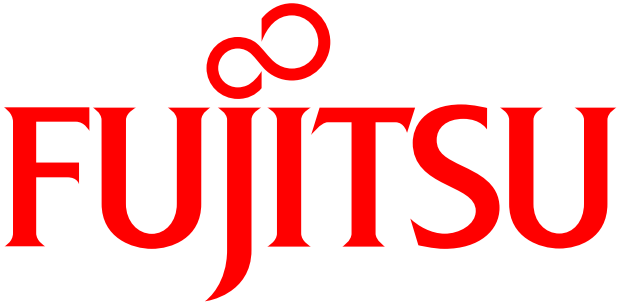
"K" Supercomputer & FX100 will provide simulation and pre-processing technology

HPC

Deep Learning

FUJITSU DLU™ Deep Learning Unit

Zinrai Deep Learning & DLU will offer a high-speed learning environment



shaping tomorrow with you