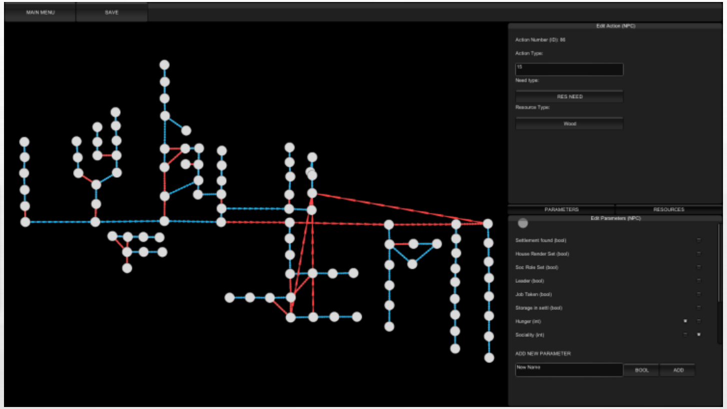
The convergence of HPC and AI

ISC2019

HPC + AI convergence

- **Scale-out of AI will be achieved through the use of HPC architecture**
  - Parallel processing – MPI
  - High speed interconnect
  - High-speed parallel storage subsystems

HPC + AI converged platform



Edit Action (NPC)

Action Number (ID): 86

Action Type: 15

Need type: RES NEED

Resource Type: Wood

Edit Parameters (NPC)

Settlement found (bool) □

House Render Set (bool) □

Soc Role Set (bool) □

Leader (bool) □

Job Taken (bool) □

Storage in settl (bool) □

Hunger (int) ■ □

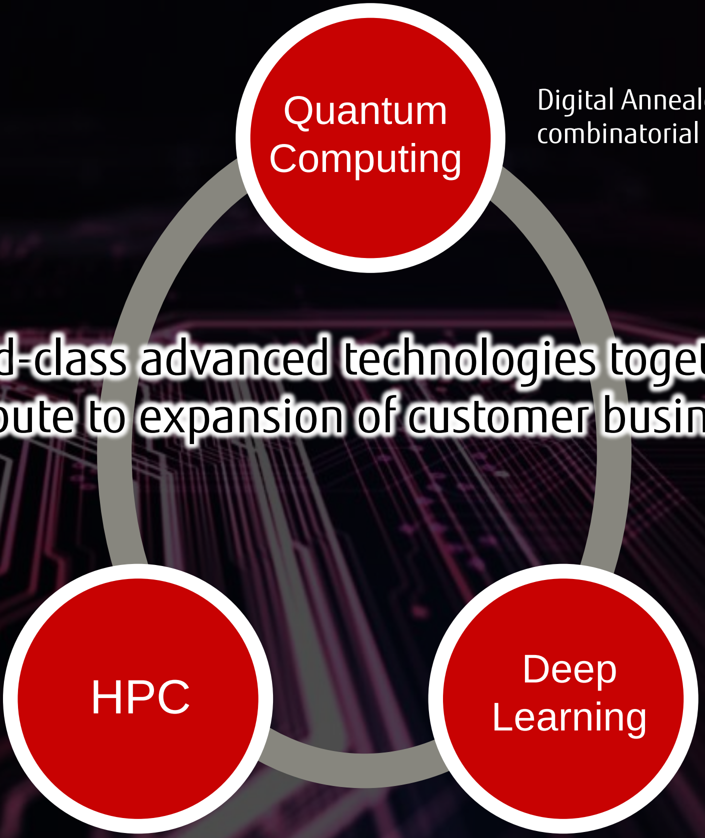
Sociality (int) □ ■

ADD NEW PARAMETER

New Name BOOL ADD

- **Certain traditional HPC problems will be solved via AI algorithms**
  - Electro-magnetics
  - Thermo dynamics
  - Computational fluid dynamics (air-flow)

AI will be accelerated by three platform technologies



Quantum Computing

HPC

Deep Learning

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

Digital Annealer will target combinatorial optimization solutions

Post-K will provide both traditional HPC as well as AI processing technology

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

« DLU » Fujitsu Deep Learning Unit

DLU - Deep Learning Unit

Processor Designed for Deep Learning

Features

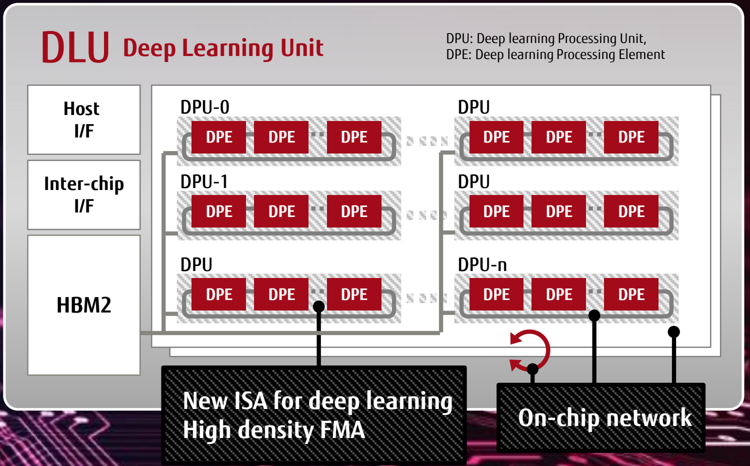
- Architecture designed for deep learning
- Low-power consumption design

  » Goal: 10x Performance / Watt compared to competitors

- Scalable design with Tofu interconnect technology

  » Ability to handle large-scale neural networks

Utilizing technologies derived from the K computer


DLU Deep Learning Unit

DPU: Deep learning Processing Unit,
DPE: Deep learning Processing Element

New ISA for deep learning
High density FMA

On-chip network

- 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

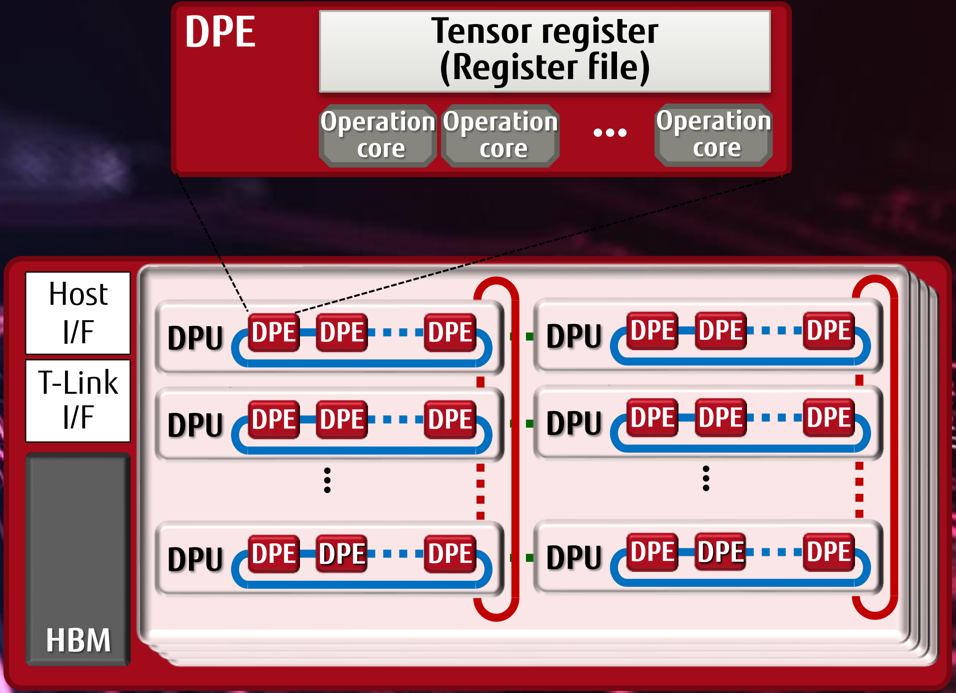
Why is DLU Fast ?

Optimized architecture for DL

- ✓ Heterogeneous core consisting of master core and operation core
- ✓ A large amount of operation cores for FP32 almost 3x larger than accelerators

DL-dedicated accuracy "DLINT"

- ✓ Realizing same accuracy of FP32 with 1/4 amount of data
- ✓ Simple integer operation contributing to low power and small chip footprint



DLU configuration

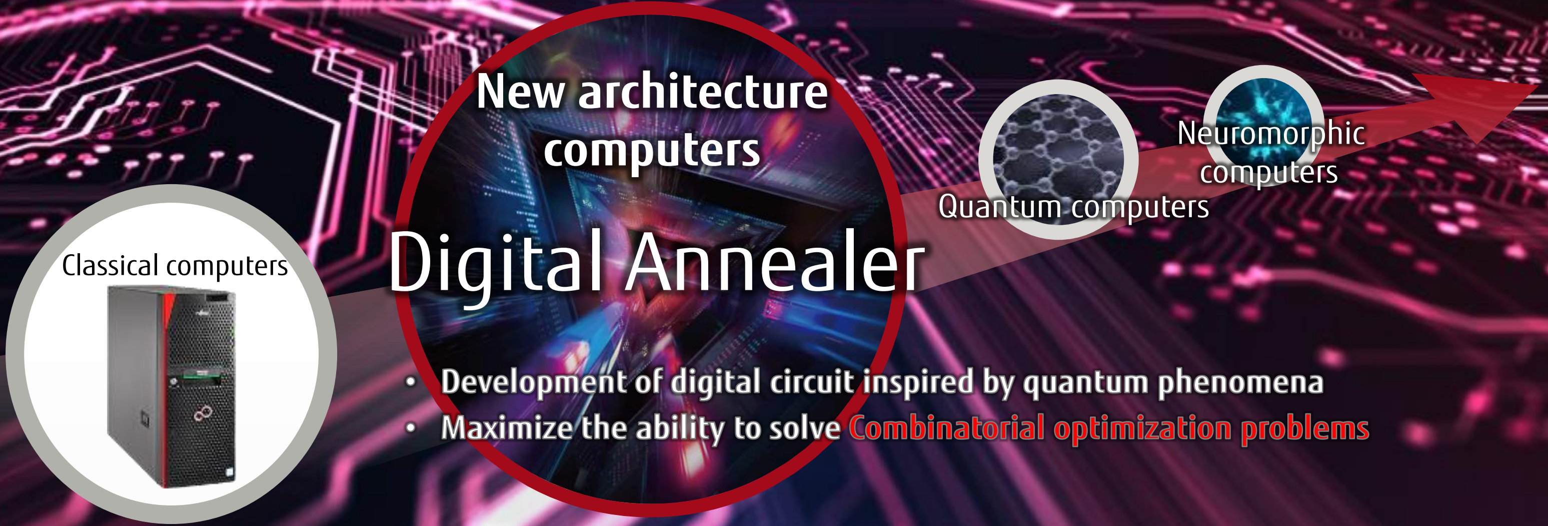
DPU: Deep Learning Processing Unit
DPE: Deep learning Processing Element

« DA » Fujitsu Digital Annealer

Digital Annealer

A new architecture to solve “Combinatorial optimization problems”

- Quantum computer still has many problems to be solved makes it difficult to apply to practical use
- The new architecture, “Digital Annealer" is to solve "combinatorial optimization problems" at high speed with digital circuit which was inspired by quantum phenomena



- Development of digital circuit inspired by quantum phenomena
- Maximize the ability to solve Combinatorial optimization problems

What Digital Annealer is...

FUJITSU

Enhanced Annealing

Digital Annealer implements in hardware, a computational technique modeled on the industrial process for tempering steel, to find a near-optimal solution in a predictable amount of time

Digital Interface

Operates at room temperature and can be installed in conventional data centers and system racks, making it far more accessible

Available As Cloud Service

Can be consumed remotely (available now), or deployed on-premises (planned), using REST API interfaces and SDKs

What it is not...

FUJITSU

AI Accelerator

Digital Annealer is not a new type of GPU

A Quantum Computer

Digital Annealer is inspired by quantum concepts; It is not a quantum device

Operationally Exotic

Digital Annealer does not require special environments or conditions to work

Big Data Engine

Relatively small memory footprint not suited to “streaming” big data applications

More effective, less dangerous cancer radiation therapy

Digital Annealer accelerates new drug and materials discovery, by finding new correlations between molecules to help develop healthier foods, prevent diseases and discover individually customized drugs

Issues

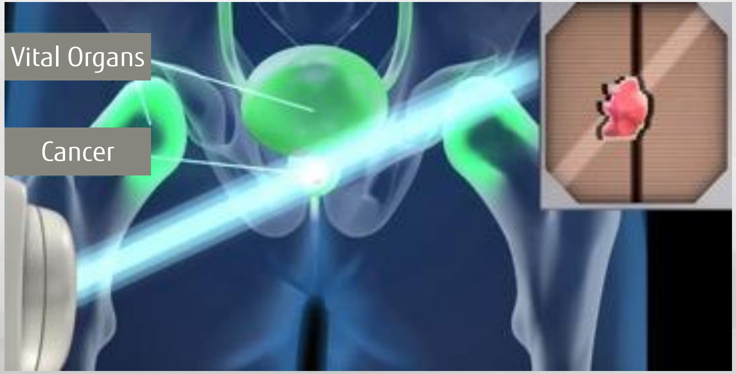
- Cancer radiation therapy can damage vital organs
- Massive number of irradiation patterns (number of combinations) with variations such as range, direction, and intensity of irradiation
- Huge computational load required for treatment plan simulation
- Even when the beams are from only one direction, the number of combinations would be 10^{150}
- Current technology needs multiple hours to a few days to calculate the combinatorial optimum

Solution

- Something here about the algorithm/method?
- Digital Annealer takes only a few minutes

Benefits

- Faster treatment plans means ability to help patients more quickly
- More accurate therapy reduces risk of side effects



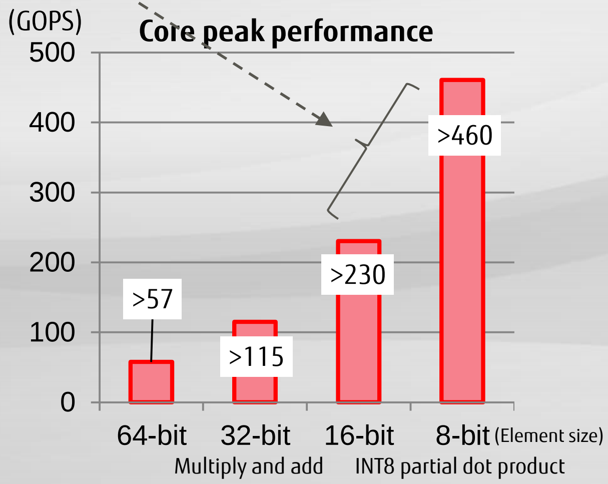
Intensity Modulated Radiation Therapy (IMRT)

In case of irradiation of 1cm^2 tumor with precision of 1mm^2 and 32 intensity levels from one direction

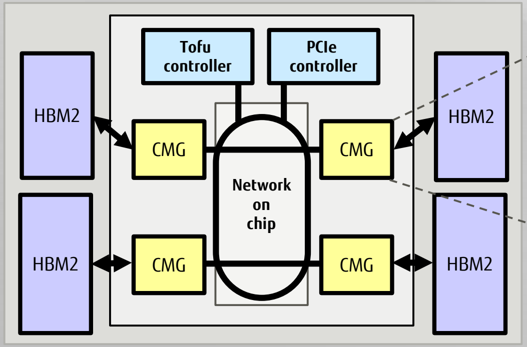
« Post-K computer » High End ARM supercomputer

Post-K computer ARM based system

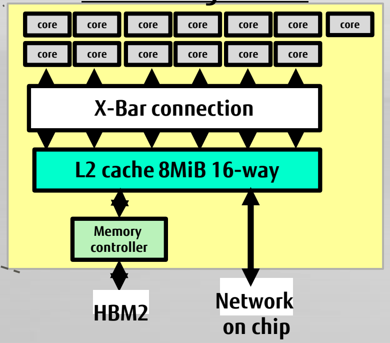
- ARM V8+SVE based processor
- ISA with **AI based instruction set** (8/16bit integer)
- High speed HBM2 memory, 6D Mesh/Torus interconnect
- High application performance and good power efficiency
- Good usability and better accessibility for users
- Keeping application compatibility while advancing from predecessors



A64FX package configuration



CMG configuration



© 2019 FUJITSU

FUJITSU

shaping tomorrow with you