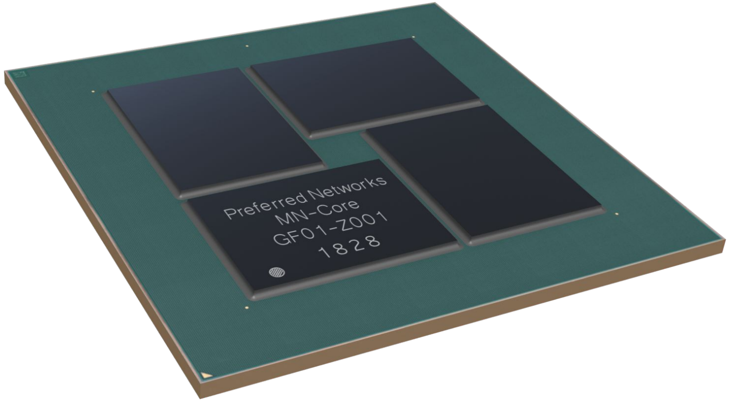
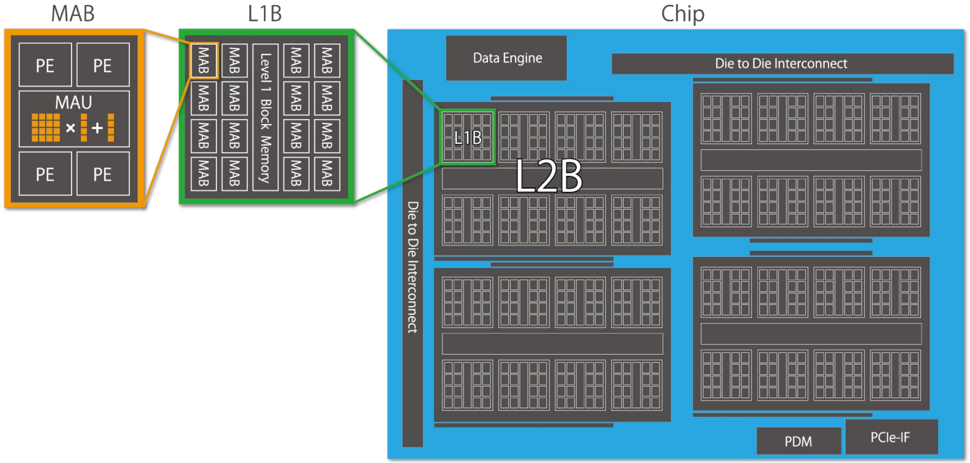
IWMLH
@ISC2020

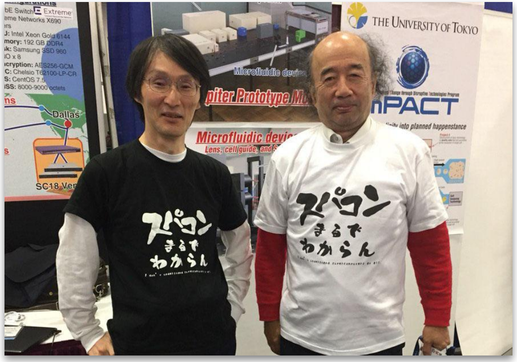
MN-Core: Massively SIMD Deep Learning Accelerator

Yusuke Doi, Ph.D
Corporate Officer, VP of Computing Infrastructure
Preferred Networks Inc.

MN-Core







In collaboration with Prof. Makino (Kobe-U) with his team members, and Prof. Hiraki (U-Tokyo, now he is with PFN),

Overview

- Introduction
  - Preferred Networks
  - MN-Core
    - Goal
- Programming Model
- Performance
- Installation
- Conclusion

Preferred Networks

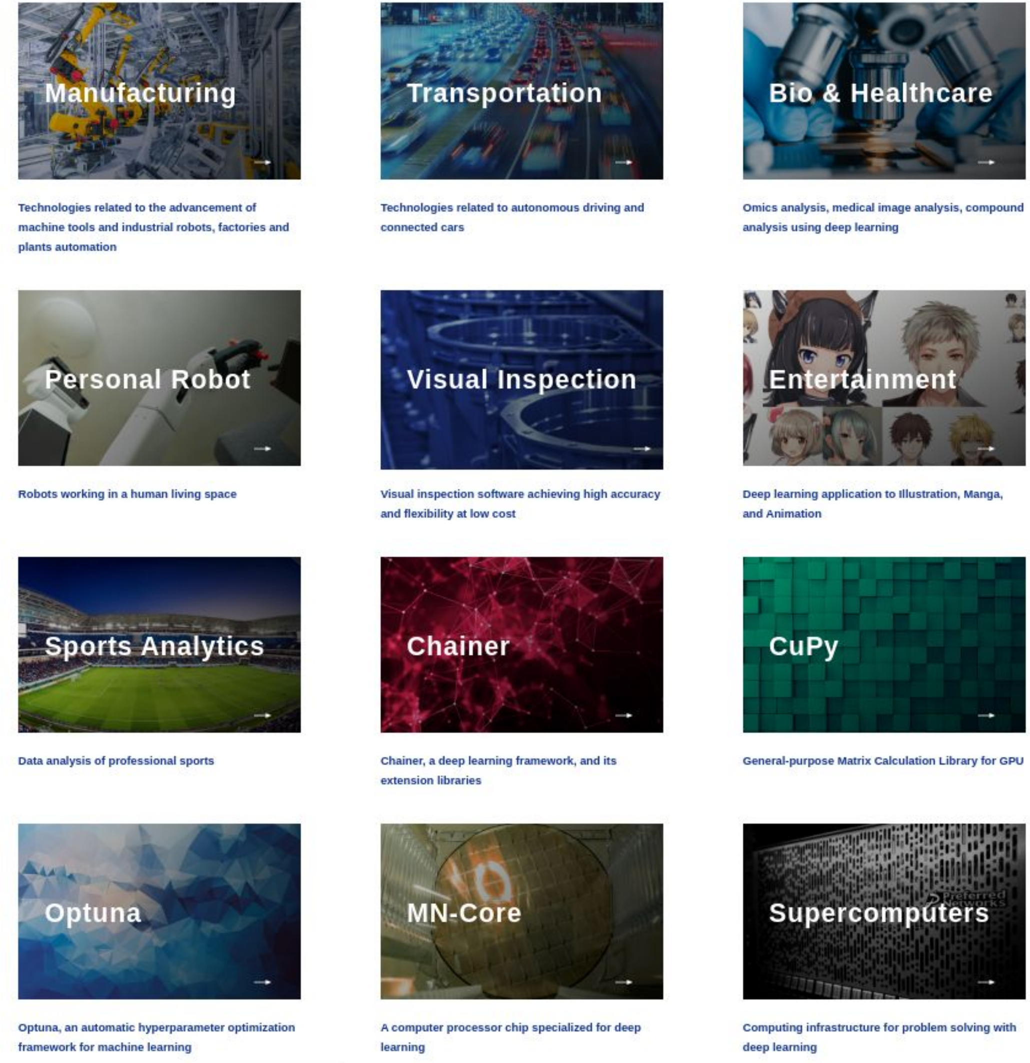
Founded

March 2014

Located

Tokyo, Japan (HQ)
Burlingame, CA., US
(Preferred Networks America, Inc.)

Make the real-world computable



Manufacturing

Technologies related to the advancement of machine tools and industrial robots, factories and plants automation

Transportation

Technologies related to autonomous driving and connected cars

Bio & Healthcare

Omics analysis, medical image analysis, compound analysis using deep learning

Personal Robot

Robots working in a human living space

Visual Inspection

Visual inspection software achieving high accuracy and flexibility at low cost

Entertainment

Deep learning application to Illustration, Manga, and Animation

Sports Analytics

Data analysis of professional sports

Chainer

Chainer, a deep learning framework, and its extension libraries

CuPy

General-purpose Matrix Calculation Library for GPU

Optuna

Optuna, an automatic hyperparameter optimization framework for machine learning

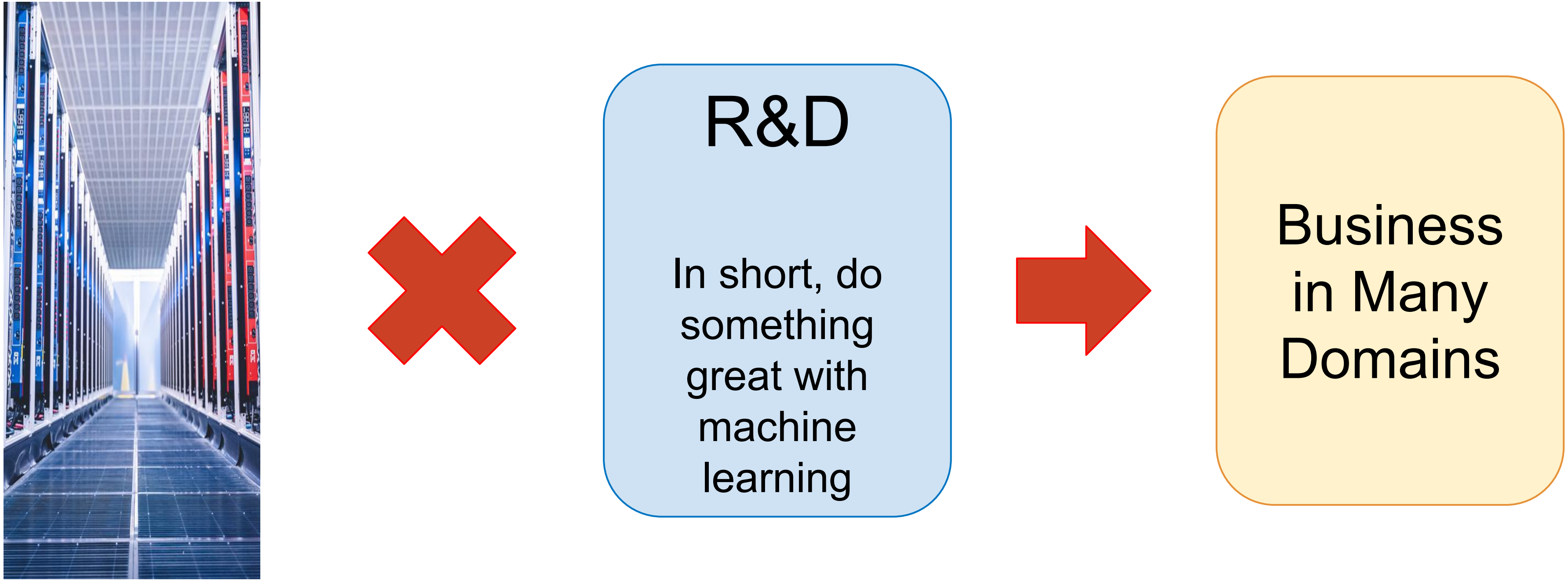
MN-Core

A computer processor chip specialized for deep learning

Supercomputers

Computing infrastructure for problem solving with deep learning

We Need Computing Power



R&D

In short, do something great with machine learning

Business in Many Domains

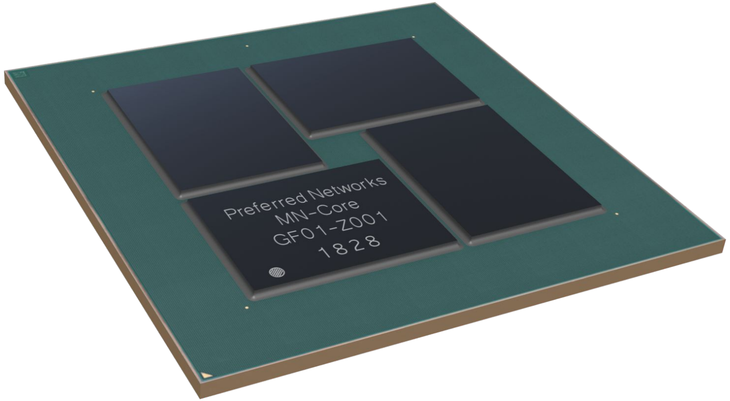
We Made What We Need: MN-Core

Accelerator for deep learning

4-die package / 500W max

Design peak performance and performance per watt:

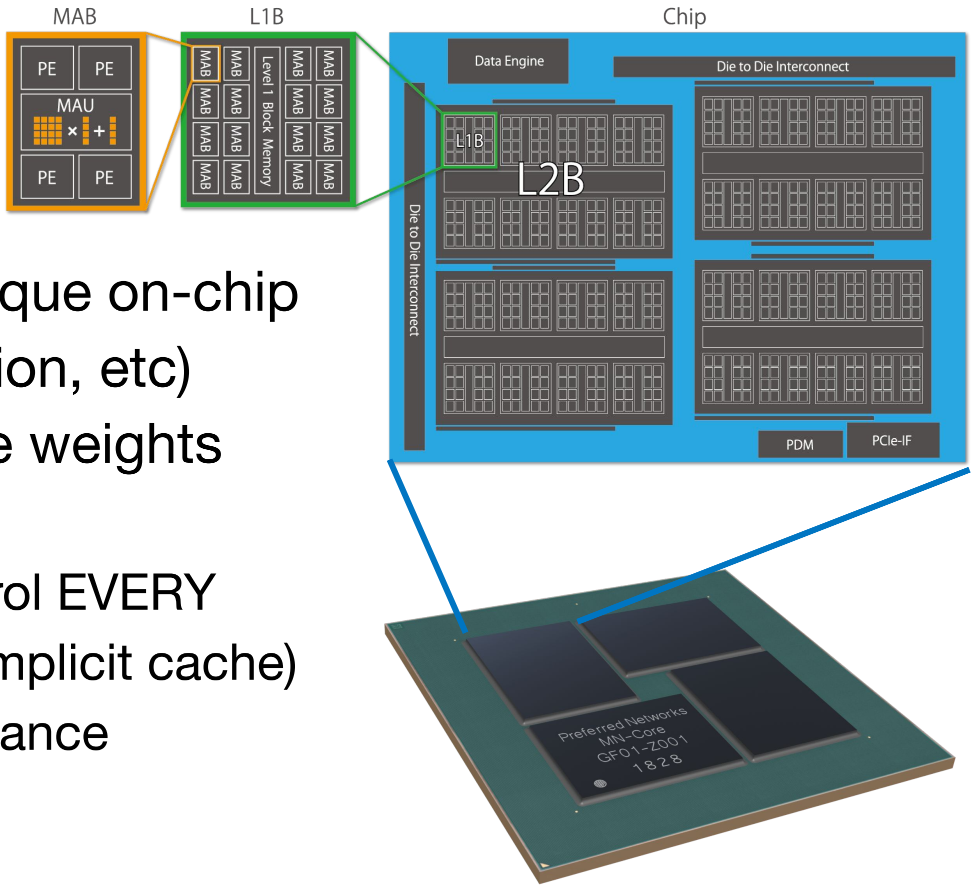
- DP: 32.8 Tops / 0.066 Tops/W
- SP: 131 Tops / 0.26 Tops/W
- HP: 524 Tops / **1 Tops/W**



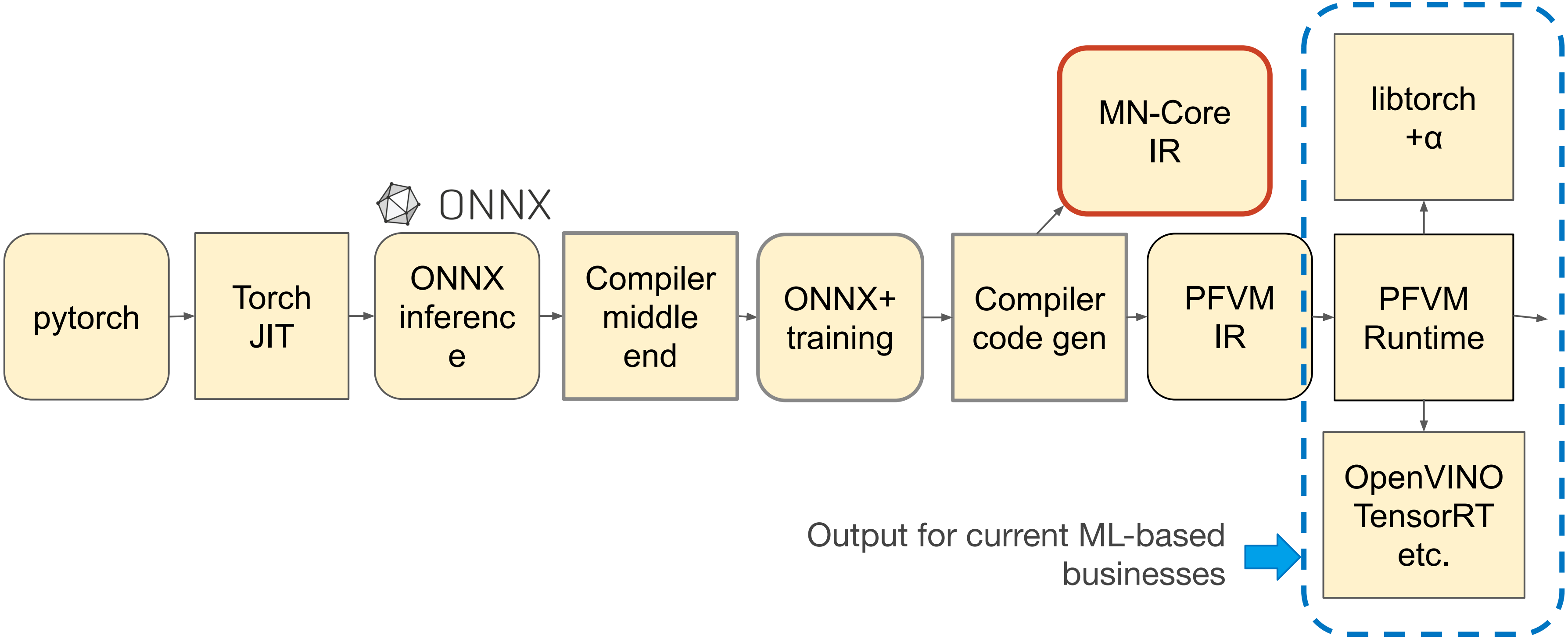
Design Overview

Giant SIMD Processor

- **Single instruction stream**
- Hierarchical structure with unique on-chip network (broadcast, aggregation, etc)
- Large SRAM to accommodate weights and filters in-place
  - programmers can/shall control EVERY memory copy explicitly (no implicit cache)
  - Easier to predict the performance



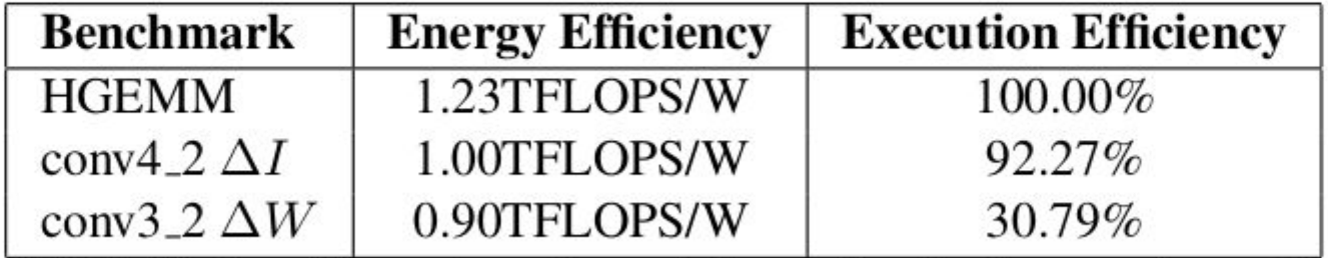
Software Strategy



Output for current ML-based businesses

- Box: software component / Rounded corner: data

Performance: Single Board



Benchmark	Energy Efficiency	Execution Efficiency
HGEMM	1.23TFLOPS/W	100.00%
conv4_2 ΔI	1.00TFLOPS/W	92.27%
conv3_2 ΔW	0.90TFLOPS/W	30.79%

Preliminary evaluation at our office →

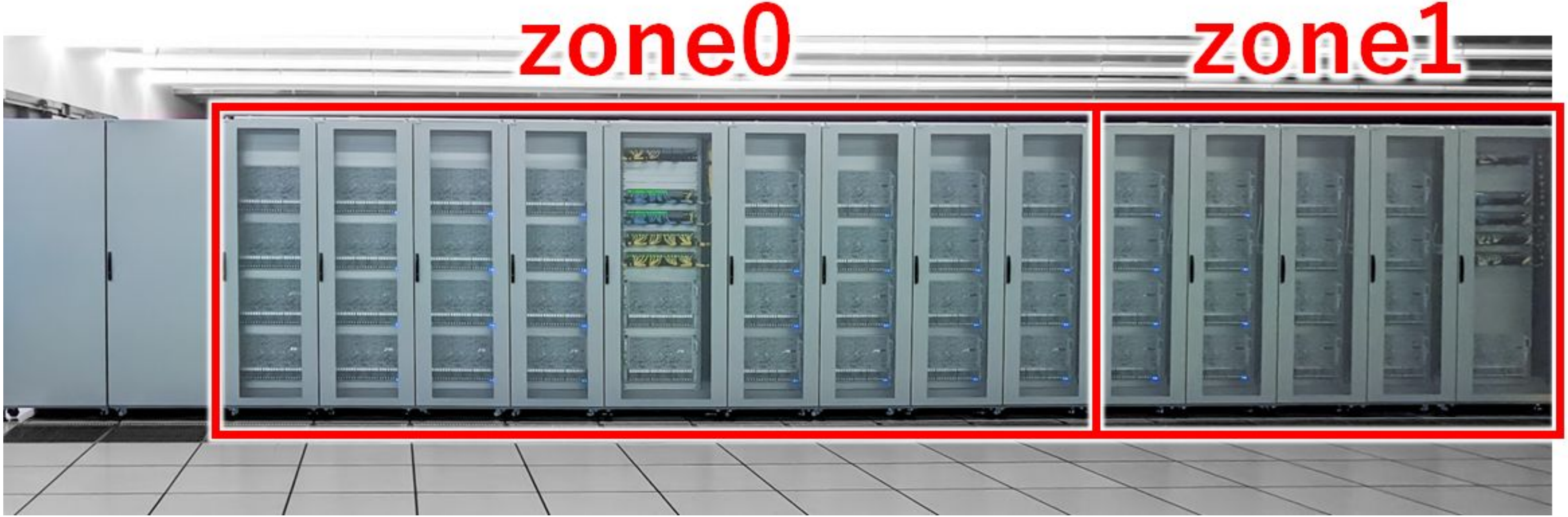
Installation: MN-3

48 computing nodes

4 x MN-Core board + 1 MN-Core DirectConnect per node

7U-high server to hold non-PCIe-compatible MN-Core board

Installation: MN-3



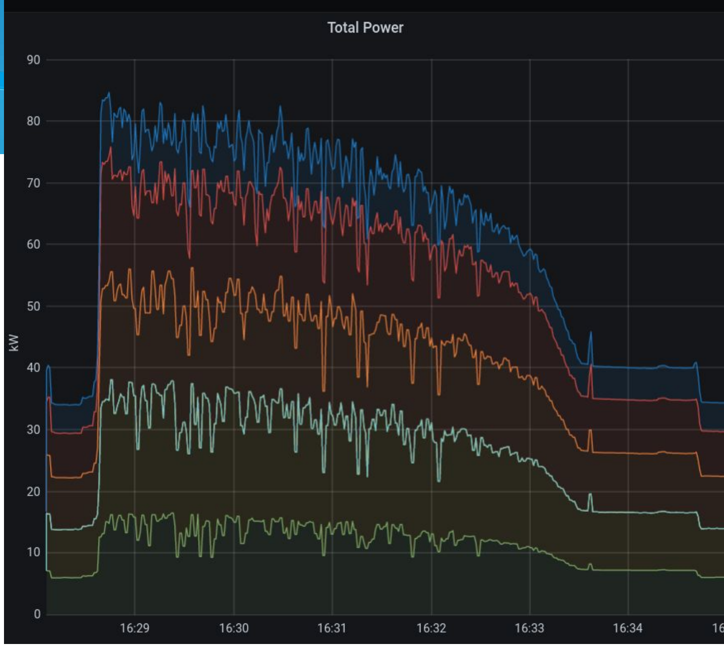
HPL Result

160 boards (40 nodes)

Power Efficiency: 21.10GFlops/W

Rmax: 1.62e+6 GFlops

Power consumption: 76808W average



Rpeak: 3.92e+6 GFlops

41% efficiency -- note that our system is designed for deep learning workload (not optimal for HPL)

Summary

- We already have our DL software eco-system and business
- Outstanding points of MN-Core
    - Huge SIMD processor, with (almost) everything in the programmer's hand
        - 524TFlops(HP)/500W, Confirmed efficiency with some hand-crafted kernels
        - The approach to give developers ultimate control
    - ONNX to IR
        - offline scheduling is the key to make this processor to work efficiently

Thank You!

Project introduction

- MN-Core: https://preferred.jp/en/projects/mn-core/
- MN-3: https://preferred.jp/en/projects/supercomputers/

If you have any question, please contact us at mncore-contact@preferred.jp