Japanese HPC processor projects

Jun Makino
Kobe University/RIKEN Center for Computational Science

8th China Computer Federation TaihuSymposium, June 22, 2019, Wuxi

Talk structure

- Introduction — comparison of processors
- Fugaku
- PEZY-SC series
- MN-Core
- Summary

The processors

- I'll discuss three ongoing Japanese projects to develop processors for HPC: Fugaku (post-K), PEZY-SCx, and MN-Core.
- Fugaku will be the first Japanese Exascale system to be completed in 2020-2021 timeframe, based on ARM Scalable Vector Extension (SVE).
- PEZY-SC2 is the current generation of the processors from PEZY computing. It has been #1 of the Green500 list since June 2017. It's based on a proprietary MIMD processor architecture with hierarchical cache. SC2 has 2048 cores and SC3 is under development.
- MN-Core is a processor developed by Preferred Networks (and JM and a few others) mainly for AI (DNN) but can be used for HPC.

Comparison

	Fugaku	PEZY-SC2	PEZY-SC3	MN-Core
Peak TF (DP/SP/HP)	2.7+/5.6/11.2	2.9/5.8/11.6	40/80/160	32.8/131/524 (4 dies/package)
Process	N7	16FFC	N7?	12FFC
Availability	2020-21	2017	2020?	2020
GF/W	15(system)	18(HPL)	80(chip)	66 (chip)
B/F	0.3	0.03	0.03	—
cores/chip	48(52)	1984(2048)	4096?	2048 (×4)
Die size(mmsq)	600?	620	700?	750×4 dies

Fugaku

- **Processor used is Fujitsu A64FX.**
- **Arm v8.2-A SVE ISA. Two 512-bit-wide SIMD units. FP16 support.**
- **48 computing and 4 assistant cores using TSMC N7 process.**
- **performance of $>$ 2.7 Tflops and 15Gflops/W.**
- **Processor up and running. System will be ready by 2020-2021.**

Let me summarize the past history of Japanese National projects for HPC systems

Japanese National Projects for Supercomputers

- MITI
  - high-speed computer system fro science and technology (1981-89)
  - 5th generation (1982-91)
  - RWCP (1992-2001)
- MEXT
  - Numerical Wind Tunnel (?-1993)
  - CP-PACS (1992-96、relatively small size)
  - Earth Simulator (-2002)
  - The K Computer (2006-2012)

Five eras of evolution of CPUs

I —1969: Before CDC7600
(before fully pipelined multiplier)

II —1989: Before Intel i860 (single-chip CPU with (almost) fully pipelined multiplier), vector-parallel

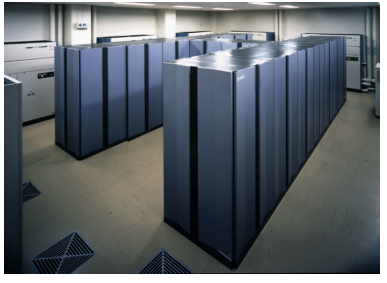
III —2003:CMOS scaling era (Power $\propto$ size^3)
From i860 to Pentium 4, scalar MPU parallel

IV —2022(?):Post-CMOS scaling era (Power $\propto$ size)
From Athlon 64 X2 to Intel MIC, multicore

V 2022(?)—: Post-Moore era(miniaturization stops)
???

Japanese National Projects for Supercomputers

- Numerical Wind Tunnel (?-1993): vector design in scalar era
- CP-PACS (1992-96 relatively small budget) typical scalar MPU
- Earth Simulator (-2002): vector design in multicore era
- The K Computer (2006-2012): early multicore design

Numerical Wind Tunnel (1993)

- Revolutionary machine which represents the transition shared memory to distributed memory. Not one-chip processor.
- Software compatibility lost. Need to develop new programs using Fujitsu-specific dialect "VPP-Fortran"
- Japanese vendors would never do this except with strong external pressure...
- Full-crossbar network.Expensive, but cheaper compared to memory interface of shared-memory machines.
- Commercial success. 10x more performance compared to shared-memory vector-parallel machines helped a lot.
- Many researchers would not mind the complete rewrite of their program, if 10x performance increase is possible.

CP-PACS (1996)

- Distributed-memory machine with one-chip microprocessors
- Pretty much a typical one-chip, one-core microprocessor.
- High B/F for a machine in mid 1990. In that sense, moving against the direction of natural evolution
- Very rich network (3-D hyper crossbar, huge amount of coaxial cables)

Earth Simulator

- 1-chip vector machine, shared memory node with 8 chips, 640 nodes in total.
- crossbar network between nodes.
- multi-chip shared memory resulted in high-cost, power-hungry memory interface
- moved backward from the NWT: from distributed memory to partially-shared memory)
- Something like the ghost of distributed

The K computer

- 8-core microprocessor, 2-way SIMD, 2-way superscalar, 4 FMA/cycle.
- Quite typical design in mid era IV.
- If we look at the details, L2 cache shared by eight cores resembles the shared-memory vector machines at the end of era II. moving backward.
- High B/F also moving backward
- multi-dimension torus network adopted 20 years after the introduction of Cray T3D. A bit outdated.

Problems within the development process of K

- There has been rumors floating around, that Fujitsu was requested to build 45-nm production line to fabricate its own processor for K.
- No matter if that rumor is true or not, the 45-nm line of Fujitsu was used only for the processor chip of K, and nothing else. The commercial version was fabricated with TSMC 40nm.
- TSMC 40nm process could pack 16 cores, while Fujitsu 45nm only 8 cores.
- It is clear that Fujitsu's investment on 45nm process was waste of money and human resources.

Accelerators for K computer?

アーキテクチャ案の概要（アクセラレータ）（6月末時点）

- 基本要求仕様
  - アクセラレータ部の理論ピーク性能：10PFLOPS
  - 汎用サーバ（ホスト）のI/Oインターフェースに接続
  - ホストより指定された演算処理をアクセラレータで実行し，結果をホストに格納

アーキテクチャ案		国立天文台	東京大学
アクセラレータ	アーキテクチャ	SIMD型 プロセッサアレイ	
	プロセッサチップ数	約 15,000	約 20,000
	ボード数	4,000	2,500
ホストサーバ数		2,000	2,500
アクセラレータ部 消費電力		-10 MW ※平成24年6月公開時の注意書き 消費電力については、10MW以下、10-20MW、20-30MWの範囲でまとめたもの。提案は、ホスト部を除いて、1.7MWであった。	-10 MW ※平成24年6月公開時の注意書き 消費電力については、10MW以下、10-20MW、20-30MWの範囲でまとめたもの。提案は、ホスト部を除いて、0.68MW（案1）、0.88MW（案2）であった。

(made public in 2012/6)

Summary of 4 proposals

	NH	F	NAOJ	UT
Peak performance (TF)	10.48	10.61	10	10
Power consumption (proposal)	23	22.8	1.7	0.88
Power consumption (MEXT document)	23	22.8	10	10

Note added in 2012

Power consumption numbers are categorized into three classes of less than 10MW, 10-20MW, 20-30 MW. The number in the proposal was 1.7MW

Why accelerators were not used?

Official explanation from MEXT

> Decided not to consider accelerators since both Fujitsu and NEC proposal meet the performance goal without using accelerators

In other words:

> Accelerators might be cheaper, or might offer better performance, but we do not consider them because our goal can be met without them.

Summary for K

- Seen as the development project:
  - Performance goal too low for the huge budget
  - Because of the outdated designs, even the low performance goal was hard. One vendor dropped out.
- The reason why the outdated designs were adopted might have been that the unspoken goal was to build "Japanese design" supercomputer, without detailed analysis of whether or not that machine will be competitive or meaningful.

Fugaku (previously known as “post-K”)

- Followed very much the same course as that of K.
- July 2011, three working groups for architecture, system software, and applications
- Official statement: to determine architecture following the requirements of applications
- JM was part of the application working group,
- JM proposed to categorize the applications in terms of required amount of memory and memory bandwidth.
- Proposed “reference” and three other types of architectures.

What happened after

- We are now at where K was in 2010.
- (unfortunately) following a very similar course
- Vector architecture has dropped out in summer 2013
- Accelerator again dropped out officially in summer 2014
- Target silently changed from 1EF to “exascale” to up to 100 times the application performance of K
- One year delay announced in 2015
- (All other exascale projects delayed anyway...)

Fugaku overview

- **Processor used is Fujitsu A64FX.**
- **Processor up and running. System will be ready by 2020-2021.**
- **Arm v8.2-A SVE ISA. Two 512-bit-wide SIMD units. FP16 support.**
- **48 computing and 4 assistant cores using TSMC N7 process.**
- **"evolutionary" design of Fujitsu HPC CPUs (HPC-2500, FX1, FX10, FX100 and A64FX)**
- **performance of > 2.7 Tflops with $B/F = 0.3$ and 15Gflops/W.**
- **B/F and performance per watt numbers are very impressive for traditional many-core CPU with wide SIMD (compare 15GF/W with Skylake or Knights Landing numbers)**

Performance Targets

- ✓ 100 times faster than K for some applications (tuning included)
- ✓ 30 to 40 MW power consumption

Peak Performance

	PostK	K
Peak DP (double precision)	400+ Pflops (34x +)	11.3 Pflops*
Peak SP (single precision)	800+ Pflops (70x +)	11.3 Pflops
Peak HP (half precision)	1600+ Pflops (141x +)	--
Total memory bandwidth	150+ PB/sec (29x +)	5,184TB/sec

* Reported in TOP500 (including I/O nodes)

Geometric Mean of Performance Speedup of the 9 Target Applications over the K-Computer

37x +

Predicted Performance of 9 Target Applications

As of 2019/05/14

Area	Priority Issue	Performance Speedup over K	Application	Brief description
Health and longevity	1. Innovative computing infrastructure for drug discovery	125x +	GENESIS	MD for proteins
	2. Personalized and preventive medicine using big data	8x +	Genomon	Genome processing (Genome alignment)
Disaster prevention and Environment	3. Integrated simulation systems induced by earthquake and tsunami	45x +	GAMERA	Earthquake simulator (FEM in unstructured & structured grid)
	4. Meteorological and global environmental prediction using big data	120x +	NICAM+ LETKF	Weather prediction system using Big data (structured grid stencil & ensemble Kalman filter)
Energy issue	5. New technologies for energy creation, conversion / storage, and use	40x +	NTChem	Molecular electronic simulation (structure calculation)
	6. Accelerated development of innovative clean energy systems	35x +	Adventure	Computational Mechanics System for Large Scale Analysis and Design (unstructured grid)
Industrial competitiveness enhancement	7. Creation of new functional devices and high-performance materials	30x +	RSDFT	Ab-initio simulation (density functional theory)
	8. Development of innovative design and production processes	25x +	FFB	Large Eddy Simulation (unstructured grid)
Basic science	9. Elucidation of the fundamental laws and evolution of the universe	25x +	LQCD	Lattice QCD simulation (structured grid Monte Carlo)

Japan's flagship machines

- ES, K and Fugaku.
- One view: High B/F, High network bandwidth, "easy to use"
- The other view: outdated design, high cost, power-hungry.

So let's now look at other Japanese processors

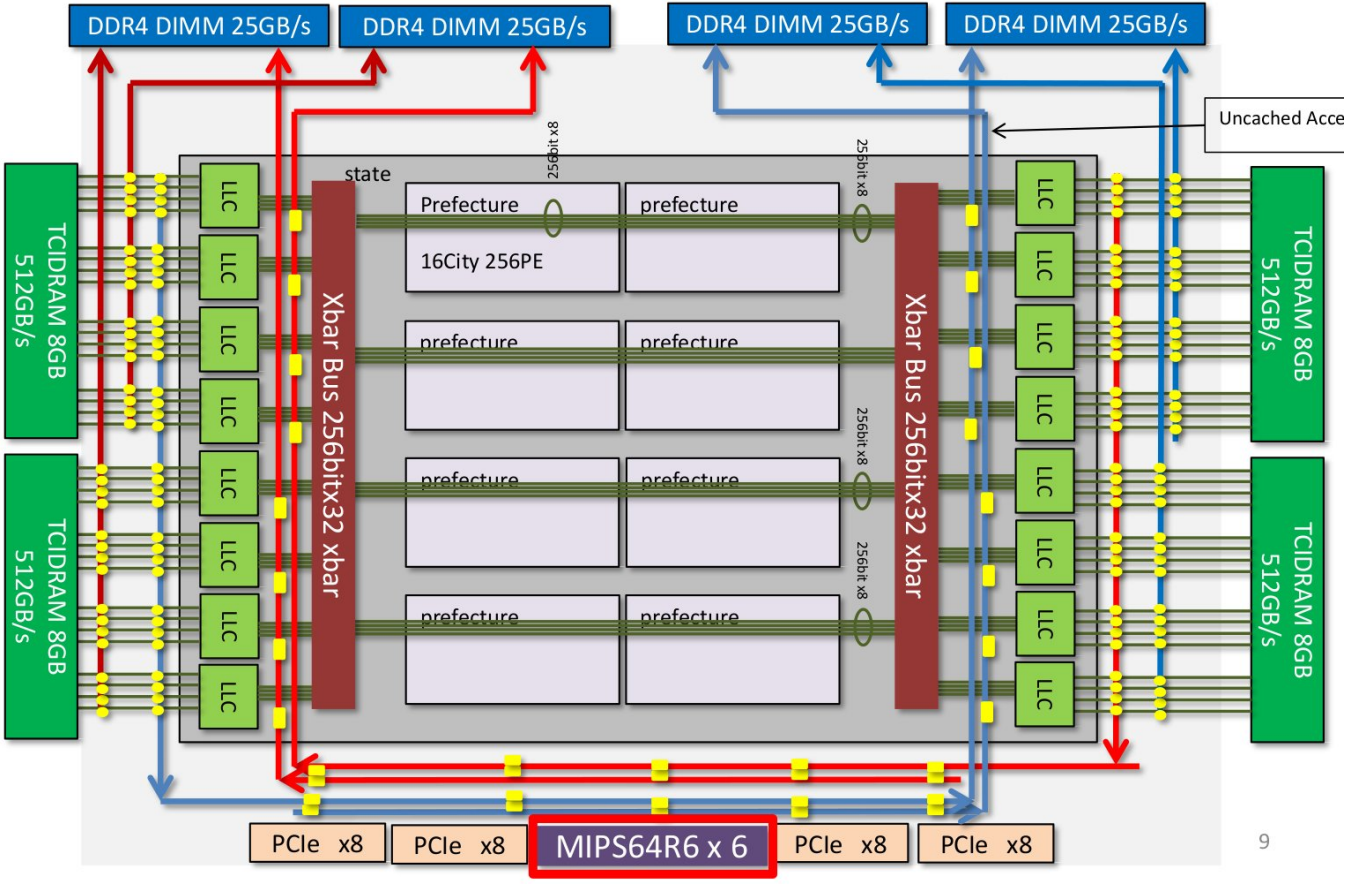
PEZY SCx

- 2 (+1 under development) generations
- 1st: PEZY-SC, 28HPM, 1024 cores, 733MHz, 1.47 Tflops. 7 HPL Gflops/W.
- Three levels of hierarchical cache. Non-coherent (explicit "flush" necessary to update L2 or lower)
- 8 DDR3/4 DRAM channels, low B/F (around 0.05).

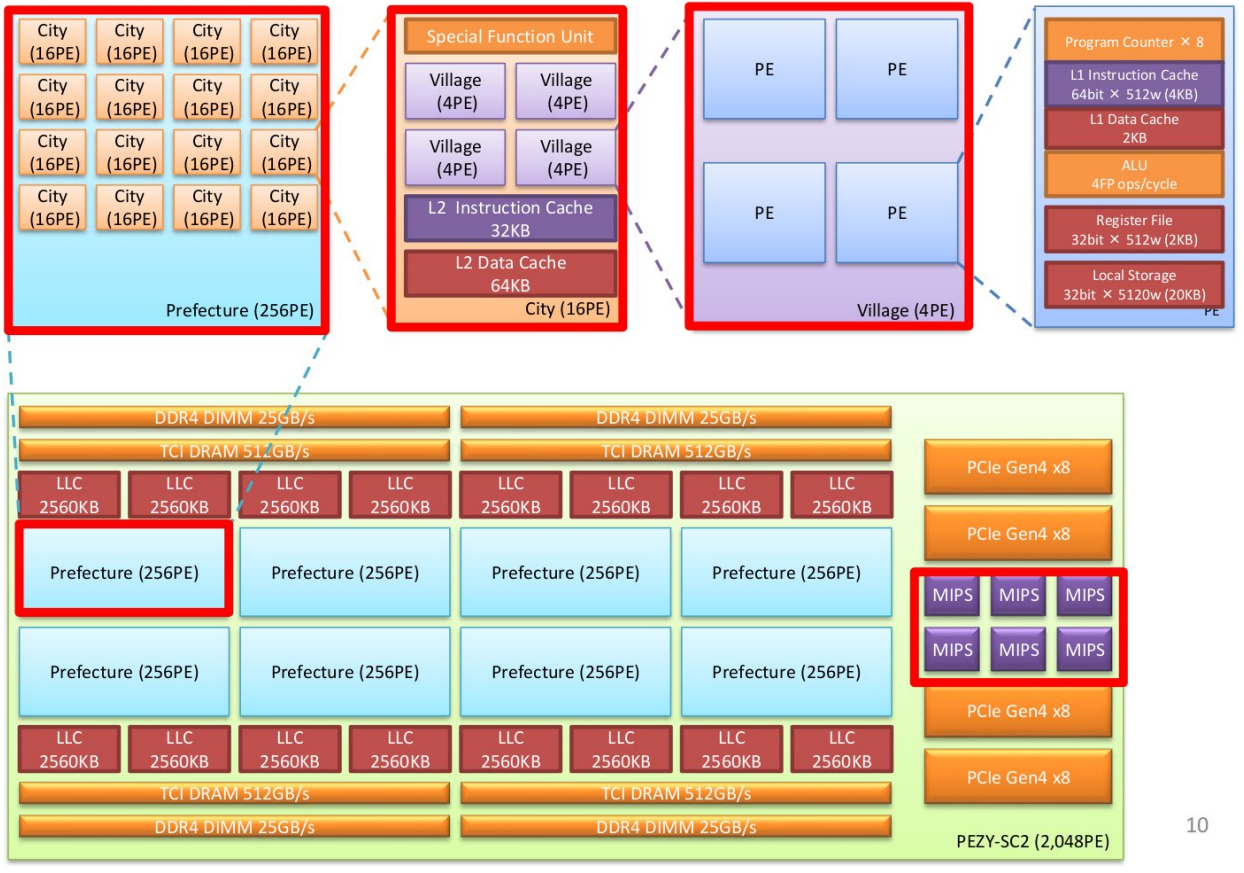
PEZY-SC/SC2 Specification

		PEZY-SC	PEZY-SC2
Process		TSMC28HPM	TSMC16FFPGL
Freq.	Core	733MHz	1GHz
	Peripherals	66MHz	66MHz
Memory	Cache	L1:1MB, L2:4MB, L3:8MB (Chip Total)	L1:12MB, L2:12MB, LLC: 40MB (Chip Total)
	Scratch Pad	16MB (16KB/PE)	40MB(20KB/PE)
IPs	Control CPU	ARM926 x 2 (Management,Debug) Cache L1:32KB x 2, L2:64KB	MIPS64R6(P6600) 6core (General Purpose)
	PCIe I/F	PCIe Gen3 8Lane 4Port (8GB/s x 4 = 32GB/s)	PCIe Gen4 8Lane 4Port (64GB/s)
	DDR I/F	DDR4 64bit 2,400MHz 8Port (19.2GB/s x 8 = 153.6GB/s)	Custom TCI Stacked DRAM 4Port 2TB/s (available on phase-2 version 2017 fall) DDR4 3.200MHz 4Port 100GB/s
Num. of PE (MIMD core)		1,024	2,048
Peak Performance		3.0T Flops (Single Precision) 1.5T Flops (Double Precision)	8.2T Flops (Single Precision) 4.1T Flops (Double Precision)
Power(typical)		70W (Leak:10W, Dynamic:60W)	130W(Estimated)

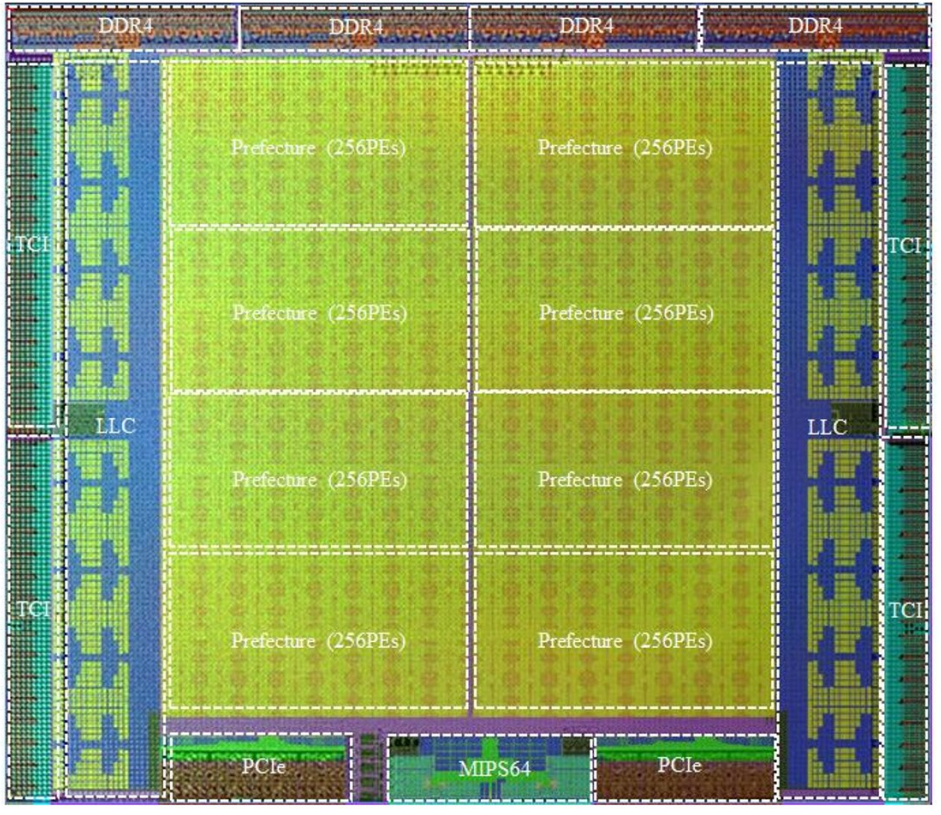
PEZY-SC2 Block Diagram



Hierarchical Architecture



Die Plot



27172.32(um) x 23695.200(um)

Software Environment

We provide OpenCL like PZCL framework
- Develop both host-processor code and PEZY-SC2 code
- LLVM is used in PZCL compiler.

Special functions for PEZY-SC2 control
- sync (barrier synchronization)
- flush (writeback from specified cache)
- get_pid, get_tid (get PEID/thread-ID)
- chgthread (change active / in-active thread)

GYOUKOU

System Overview

26 Tank 832node system (32node / tank)

Model	ZettaScaler-2.0
Nodes	832
Vendor	ExaScaler Inc.
Processor	Xeon D -1571
Speed	1,300
Sockets per Node:	1
Cores per Socket:	16
Accelerator/CP:	PEZY-SC2
Accelerators/CP per Node:	16
Cores per Accelerators/CP:	2,048
Operating System:	Linux CentOS 7.3
Primary Interconnect:	InfiniBand EDR
Memory per Node (GB)	1,088

Immersion Cooling Tanks

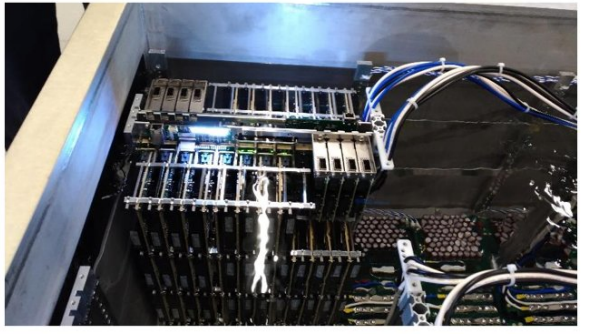
26 Tanks @ JAMSTEC
16 Bricks / Tank

Brick

Brick

1brick = 2node 32 x PEZY-SC2

Ultimate High Density Implementation

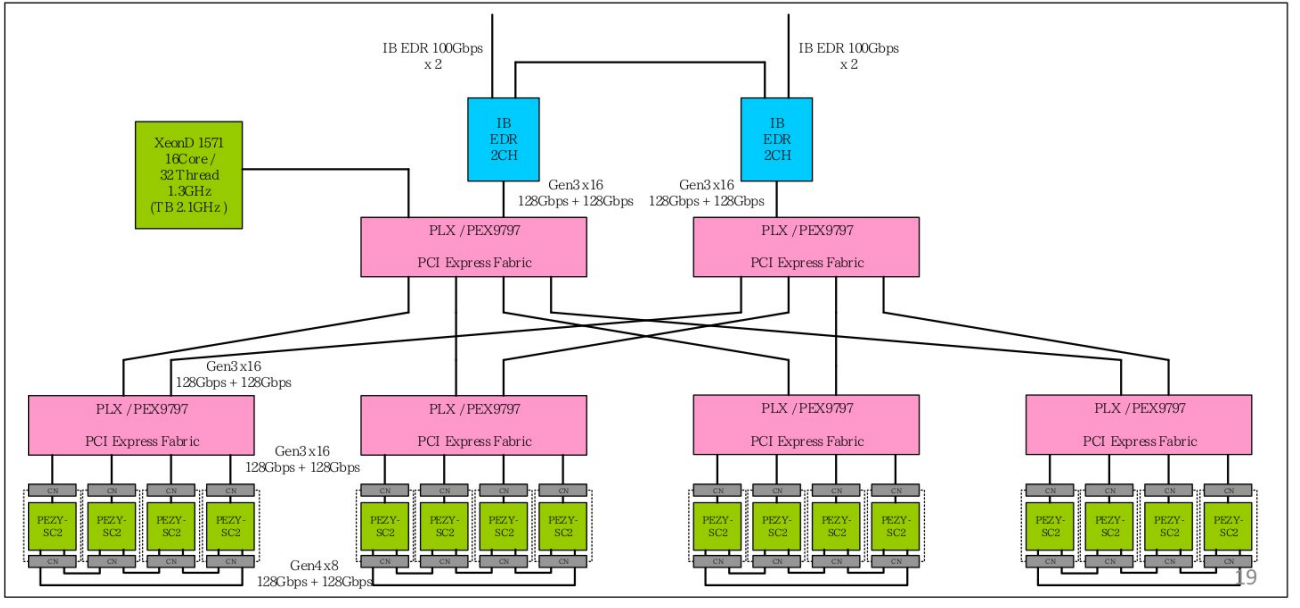


1 Base Carrier Board
8 Sub Carrier Board
32 PEZY-SC2 Module Card
1 Dual-XeonD Module Card
4 InfiniBand EDR HCA

Node Overview

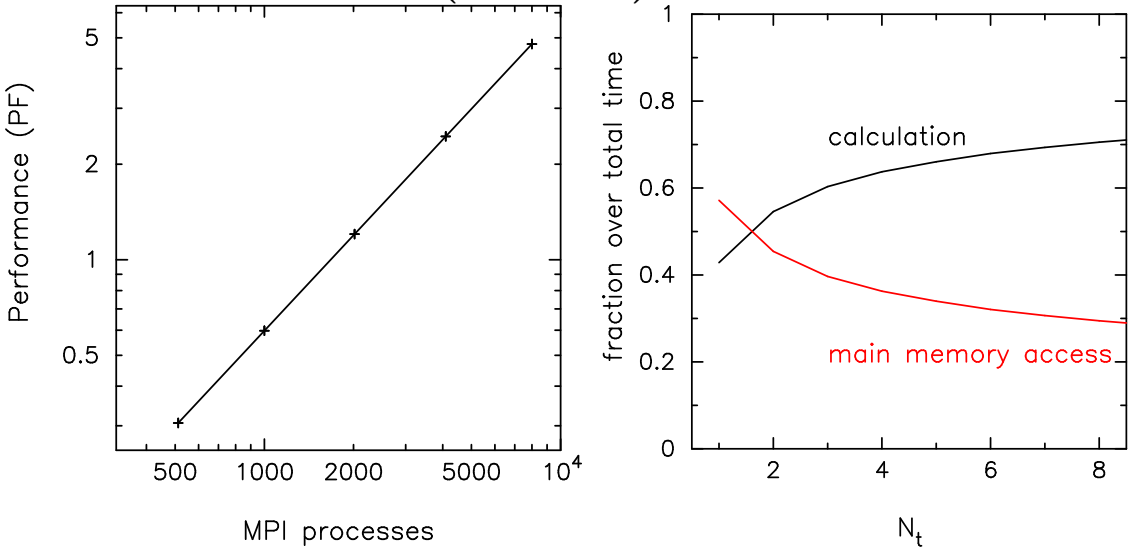
Node

1 x Xeon D-1571 (16core, 1.3GHz)
16 x PEZY-SC2 (2,048core, 1GHz)
Multi-Layer PCIe Internal Network (Gen3 x16, 128Gbps+128Gbps)
Inter-SC2 Ring Network (PCIe Gen4 x 8 128Gbps + 128Gbps)
2 x InfiniBand Inter-node Network (EDR 100Gbps)



PEZY-SC application performance (1)

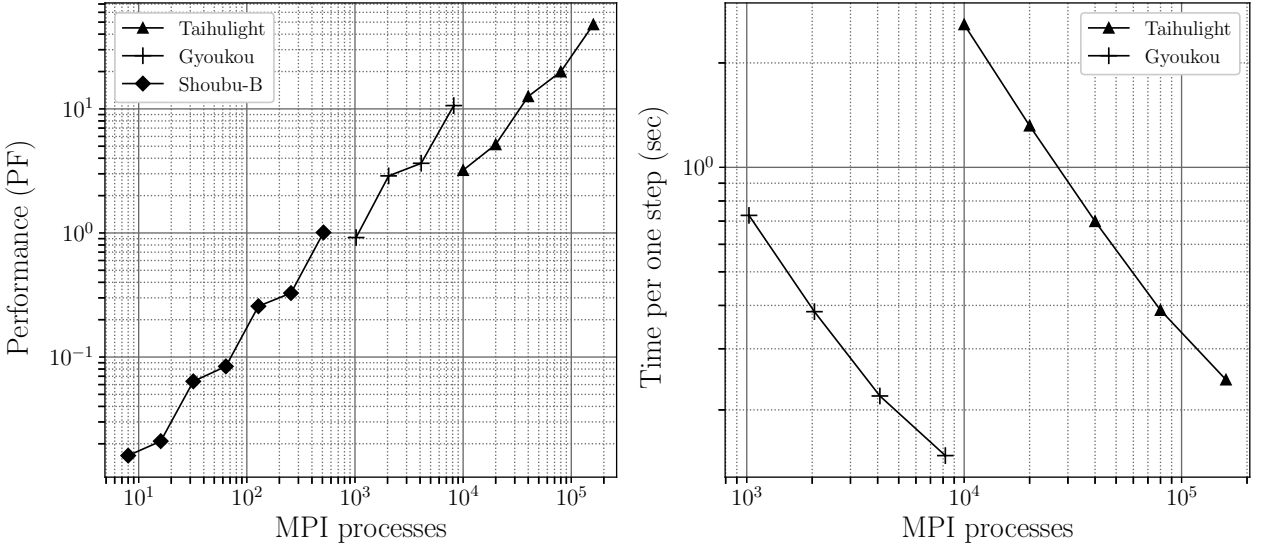
Tanaka et al. 2018 (ESPM2)



Explicit compressive CFD on regular grid with temporal blocking.
$\sim$ 20% of theoretical peak.

PEZY-SC application performance (2)

Iwasawa et al. 2019 (preprint)
N-body simulation of Saturn's rings.



If you can use TaihuLight, you may be able to use PEZY-SCx.
40 – 50% of theoretical peak.
(number of particles 10x larger for Taihulight)

President of PEZY Computing Arrested, Charged with Fraud

Michael Feldman | December 6, 2017 10:52 CET

The president of PEZY Computing, Motoaki Saito, was arrested Tuesday for allegedly defrauding the government of 431 million yen ($3.8 million). Daisuke Suzuki, another PEZY employee, was also arrested.

Among other products, PEZY supplies high performance computing processors, the most recent offering being the PEZY-SC2. The seven-year old company's capital worth is currently about 940 million yen ($8.4 million).

The PEZY-SC2 chip powers the 19.4-petaflop Gyoukou supercomputer, which is currently the fourth fastest system in the world, according to the latest TOP500 list. Besides Gyoukou, PEZY-SC2 is also installed in three other new HPC machines of note: Shoubu system B, Suiren2, and Sakura. Although not as powerful as Gyoukou, these three systems captured the top spots on the Green500 list, thanks to their superior energy efficiency.

Current status of PEZY-SC2/SC3

- **GYOUKOU was dismantled (as of June 2018)**
- **SC3 design has been completed.**
- **SC3 production schedule remains unclear.**
- **Could have been the fastest path to exaflops...**

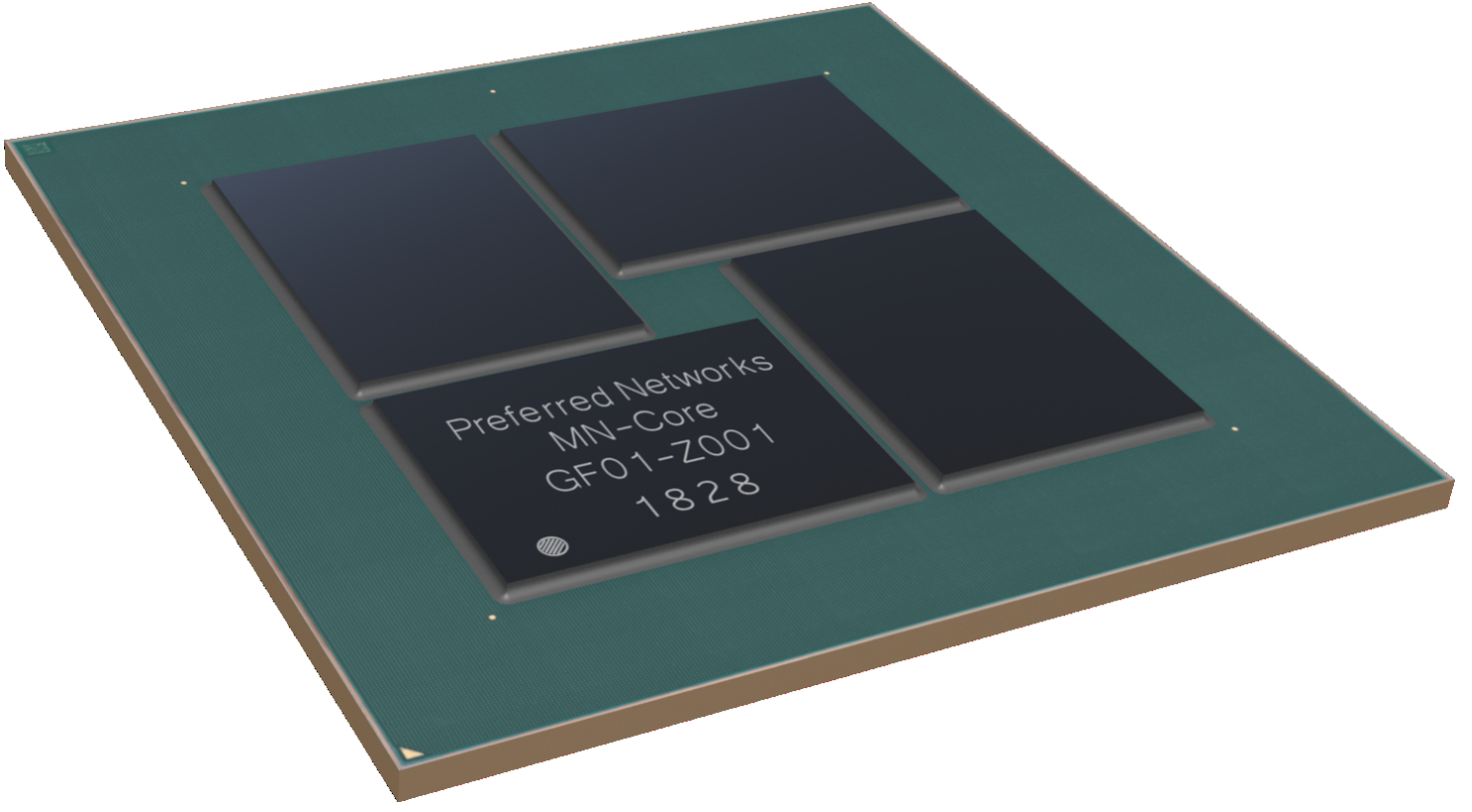
MN-Core (aka GRAPE-PFN2)

- **The processor chip PFN (Preferred Networks, a Japanese AI venture) has been developing in collaboration with some of our group in RIKEN R-CCS/Kobe University.**
- **Goal: Highest performance and highest performance-per-watt for training DNNs (CNNs).**
- **Planned peak FP16(-equivalent) performance of single card: 524 Tops**
- **Target power consumption: <500W, > 1Tops/W**

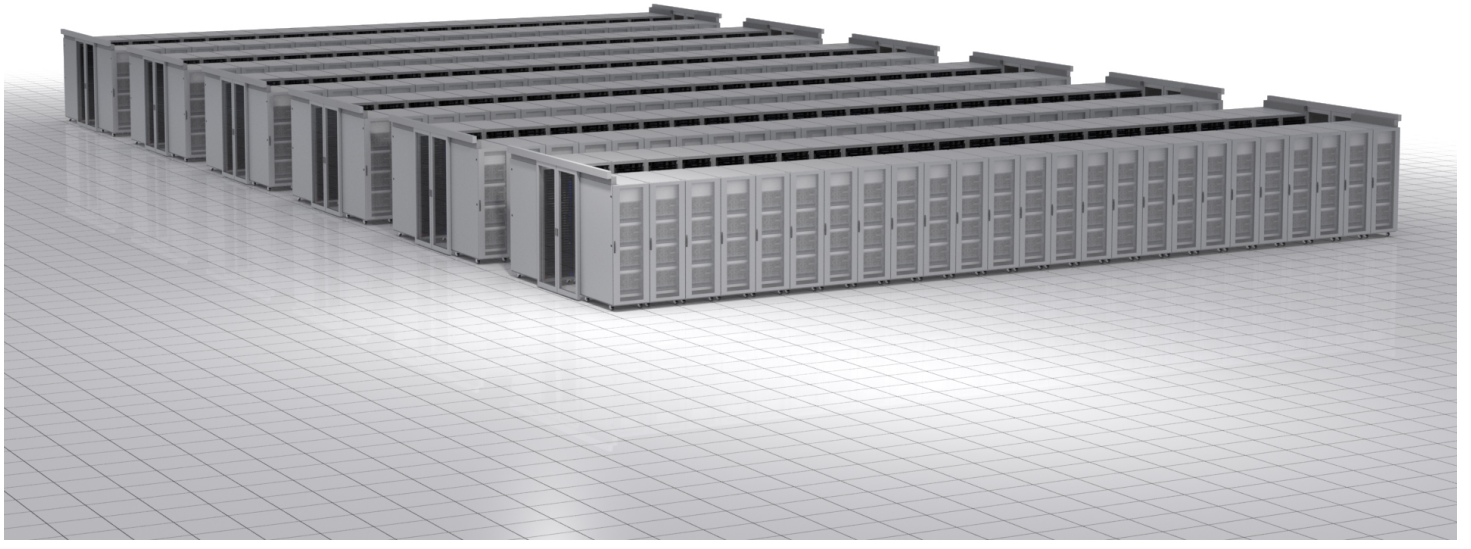
Past history, current status, and future plan.

- **Feb 2016: JM visited PFN at Hongo-3choume**
- **June 2016: Joint application to NEDO ("small" grant, 40MJYE/year $\times 2$) (PFN moved to Ote-machi)**
- **July 2016: PFN chip project started. Plan for two chips: GPFN1 by NEDO money (40nm, small chip), GPFN2 (12FFC, full-blown) by PFN internal money.**
- **July 2017: final "Go" for GPFN2 chip.**
- **2019 Evaluation of ES chips will be ...**
- **2020 "2EF" system (MN-3) will be ready at JAMSTEC ES site.**

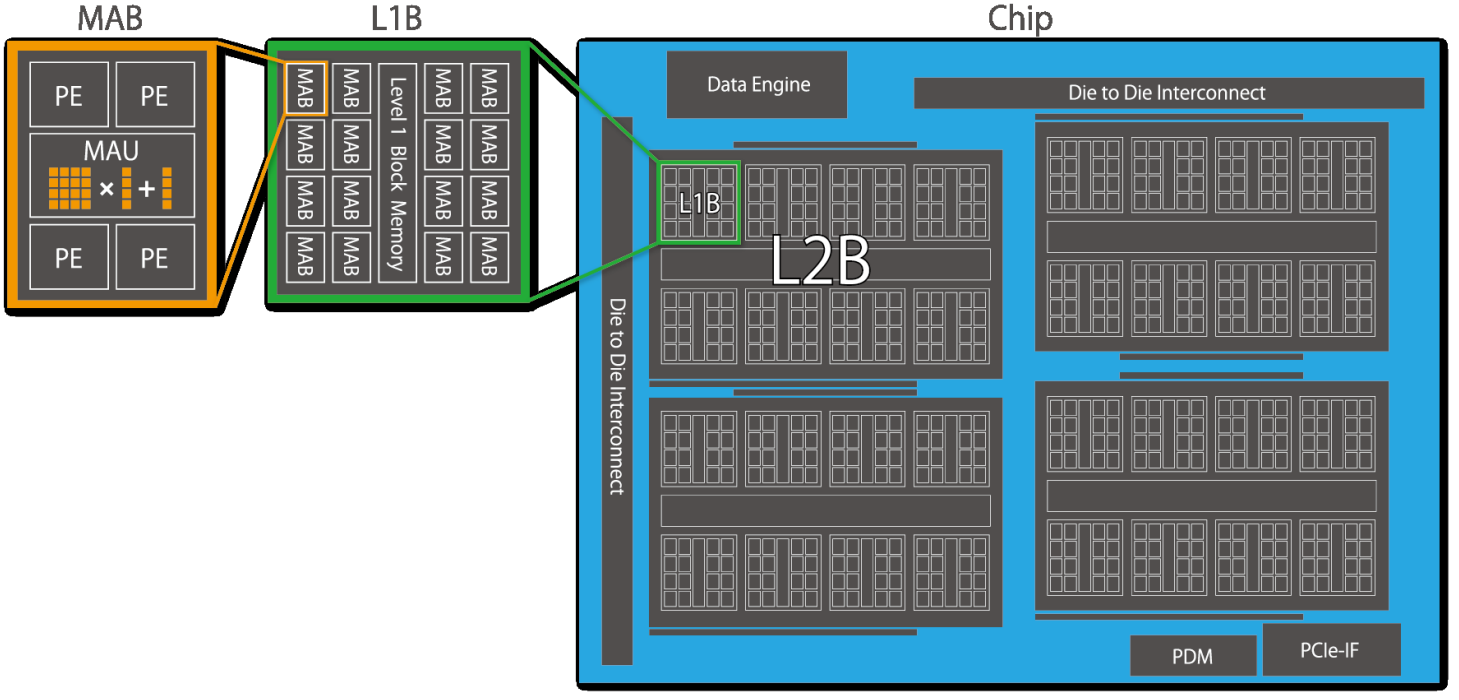
MN-Core



MN-3



GRAPE-PFN2 architecture



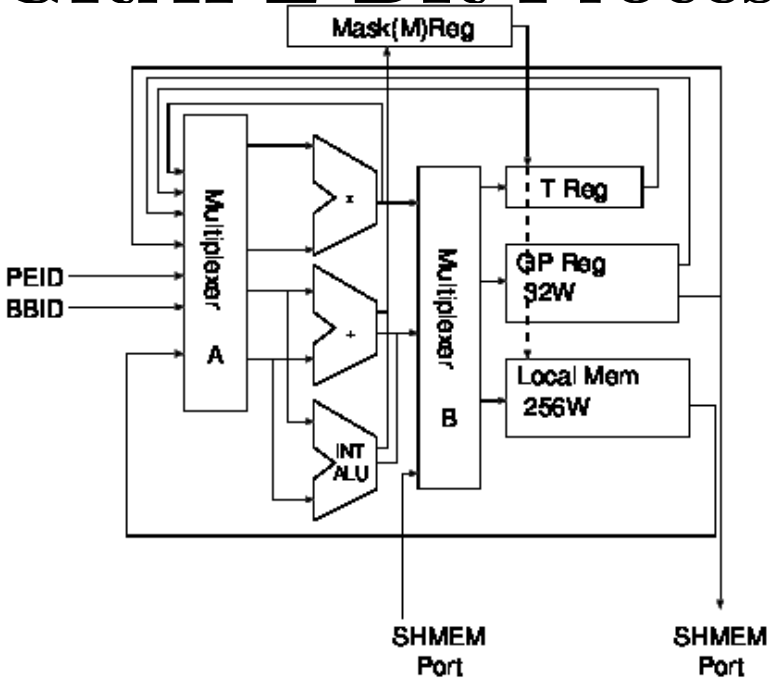
Overview of GPFN2

- One card: One "module", One module: four chips in one package
- One chip: one PCIe interface, DRAM interfaces, four "Level-2 broadcast blocks" (L2Bs)
- One L2B: eight L1Bs
- One L1B: 16 MABs (Matrix Arithmetic Blocks)
- One MAB: four Processor Elements combined to perform FP64, FP32, or FP16 matrix-vector multiplication.
- One PE can be also used as 64-bit scalar processor. We added many special instructions for DL.
- All PE/MAB/L1B/L2B operate on single clock and single instruction stream (card-level SIMD)

GRAPE-DR

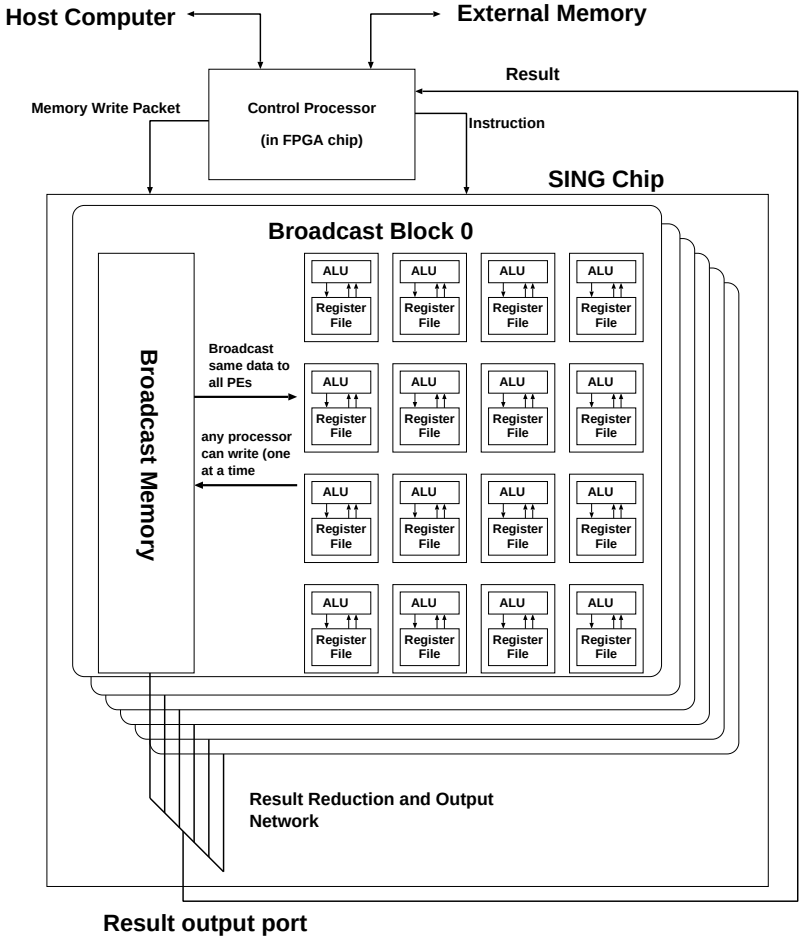
- **Many of the technical details of GRAPE-PFN2 is still undisclosed.**
- **GRAPE-PFN2 is very much a natural extension of GRAPE-DR.**
- **GRAPE-DR project started in 2004 and the machine completed in 2009.**
- **GRAPE-DR chip: completed in 2006, TSMC 90nm, 500MHz, 256 DP Gflops, $\sim$ 4 GF/W.**

GRAPE-DR Processor architecture



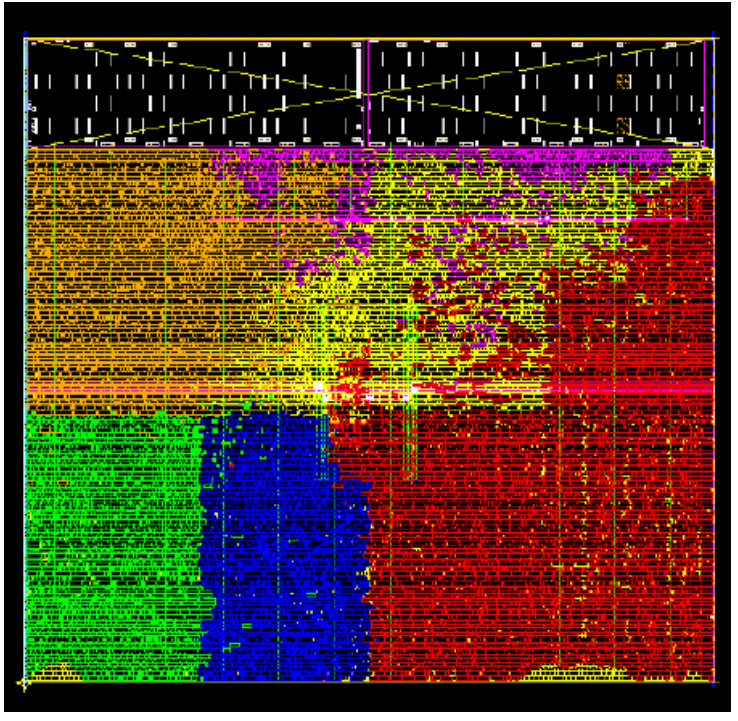
- Float Mult
- Float add/sub
- Integer ALU
- 32-word registers
- 256-word memory
- communication port

Chip architecture



- 32 PEs organized to "broadcast block" (BB)
- BB has shared memory.
- Input data is broadcasted to all BBs.
- Outputs from BBs go through reduction network (sum etc)

PE Layout

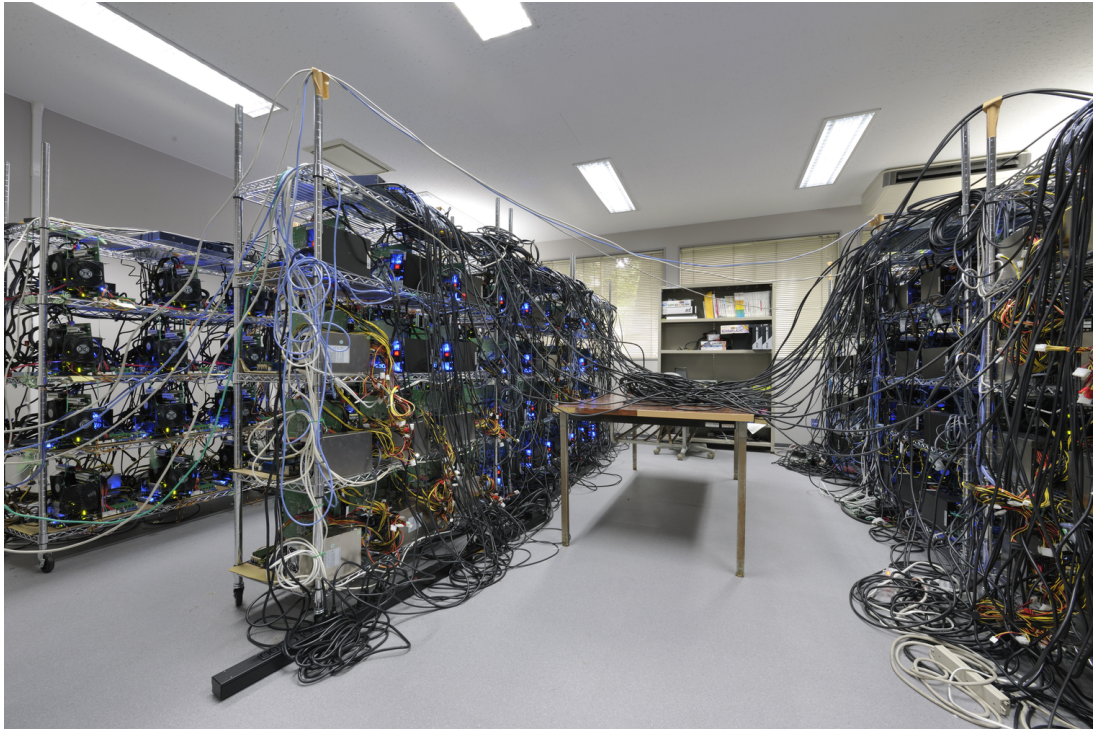


Black: Local Memory
Red: Reg. File
Orange: FMUL
Green: FADD
Blue: IALU

0.7mm by 0.7mm
800K transistors

0.1W@400MHz
800Mflops/400Mflops
peak (SP/DP)

GRAPE-DR cluster system



(As far as I know) Only processor designed in academia listed in Top500 in the last 20 years.

Little Green 500, June 2010

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	815.43	National Astronomical Observatory of Japan	GRAPE-DR accelerator Cluster, Infiniband	28.67
2	773.38	Forschungszentrum Juelich (FZJ)	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
2	773.38	Universitaet Regensburg	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
2	773.38	Universitaet Wuppertal	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
5	536.24	Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw	BladeCenter QS22 Cluster, PowerXCell 8i 4.0 Ghz, Infiniband	34.63

#1: GRAPE-DR,
#2: QPACE: German QCD machine
#9: NVIDIA Fermi

Changes made from GRAPE-DR

- Second layer of on-chip tree network
- Integration of PCIe and DRAM interface
- Addition of MAB
- Much larger memory
- Many other changes in on-cip network
- Design optimized to DNN/CNN
  (both inference and learning)

Some numbers

- **Number of MABs and PEs: 512 (2048) per chip, 2048 (8192) per module.**
- **32.8TF (DP), 132TF (SP) and 524Tops (HP)**
- **Memory bandwidth: (not yet open)**
- **Link to the host PC: PCIe**
- **On-board DRAM: 32GB**

GPFN3 project goal

- **Minimum goal: mass production cost $\ll$ 1M JYE per HP PF.**
- **Ideally including NRE.**
- **Nvidia Volta: $\sim$ 8M JYE/HP PF.**

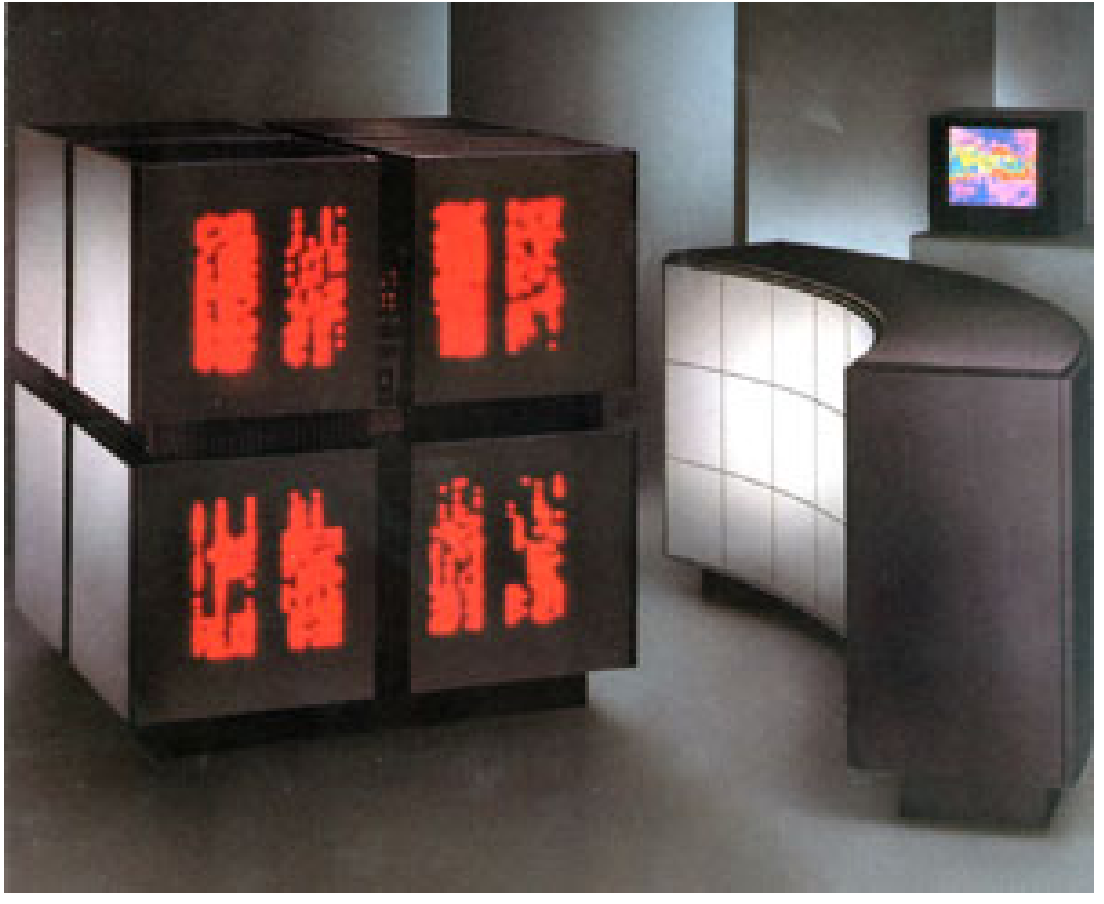
How can we achieve this goal?

- put more FPUs per die area
  - reduce memory per Processing element
  - "optimize" accuracy
- increase the clock frequency (high-clock design)
- use cheaper transistors (7nm? 12? or 22?)

HPC on MN-Core/GRAPE-PFNx

- **Support for DP/SP operations (1/16 and 1/4 speed of HP)**
- **Chip-level SIMD with local memory and on-chip network = The Connection Machine on a chip**

TMC CM-2



2048 Weitek FPUs operate as one SIMD computer

Fairly good programming environment (C*, CM-Fortran)

Summary for MN-Core/GRAPE-PFNx

- GPFN2 will be ready “soon”
- 512 Tops/132TF/32TF for half, short and long word
- GPFN2 will offer significantly better price performance compared to what is currently available in the market.
- GPFN3 goal is to improve the price performance significantly.
- they are designed HPC applications in mind.

Comparison

	Fugaku	PEZY-SC2	PEZY-SC3	MN-Core
Peak TF (DP/SP/HP)	2.7+/5.6/11.2	2.9/5.8/11.6	40/80/160	32.8/131/524 (4 dies/package)
Process	N7	16FFC	N7?	12FFC
Availability	2020-21	2017	2020?	2020
GF/W	15(system)	18(HPL)	80(chip)	66 (chip)
B/F	0.3	0.03	0.03	—
cores/chip	48(52)	1984(2048)	4096?	2048 (×4)
Die size(mmsq)	600?	620	700?	750×4 dies

Conclusions?

- Japanese Flagship projects: aiming at "ease of use" (High B/F, rich network)
- Fujitsu has a different goal (Track record: IBM $\rightarrow$ SPARC $\rightarrow$ ARM)
- Recent Japanese HPC/AI processor projects: aiming at better watt- and cost-performance
- (In my opinion, in the Post-Moore era, what matters are watt- and cost-performance)
- Now is certainly an exciting time.
- I'm very interested in how they are compared with Chinese exascale machines.