IEEE MAGNETICS

IEEE Distinguished Lecturer for 2023

Computational I/O Stack Workshop, August 17, 2023

TOHOKU UNIVERSITY

Magnetic Data Storage Technology
from the invention of perpendicular magnetic recording (PMR) to computational storage

Yoichiro Tanaka, PhD

Research Institute of Electrical Communication

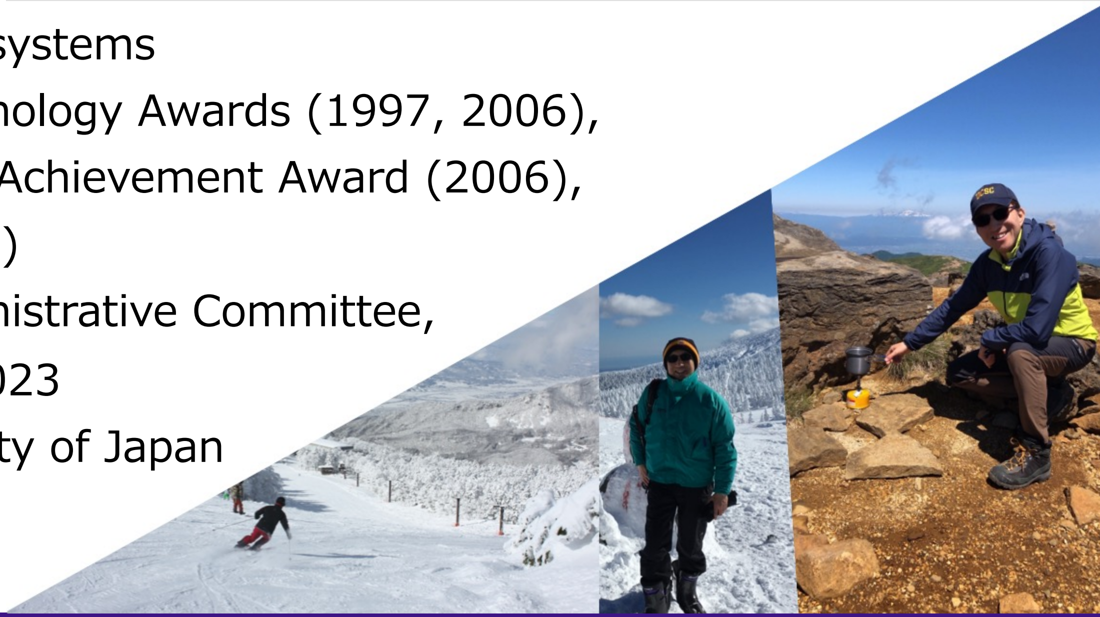
Tohoku University

yoichiro.tanaka.e1@tohoku.ac.jp

Yoichiro Tanaka, PhD

Professor
Research Institute of Electrical Communication
Tohoku University, Sendai, Japan

- Research & development of **perpendicular megnetic recording (PMR)** in industry and academia over 30 years
  - ✓ Achieved **the world's first PMR HDD to commercialize**
  - ✓ Recording physics, a giant magnetoresistive head, glanular thin film media materials, signal processing
  - ✓ Computational storage systems
- Awards: The Nikkei BP Technology Awards (1997, 2006), The Japan Magnetic Society Achievement Award (2006), Okochi Memorial Prize (2007)
- IEEE Magnetic Society Administrative Committee, Distinguished Lecturer for 2023
- Fellow of the Magnetic Society of Japan
- The Secretary General for INTERMAG 2023 Sendai

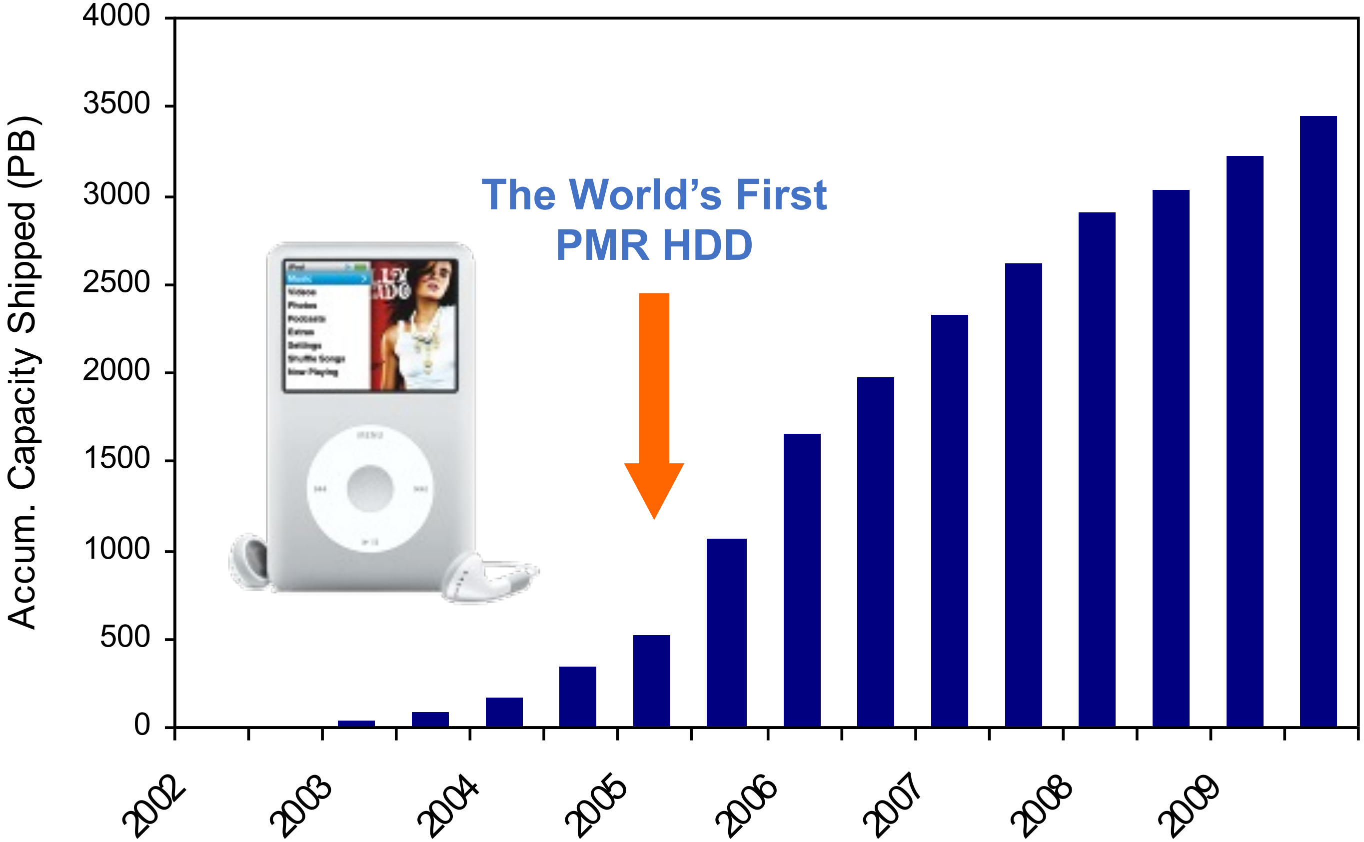


Agenda

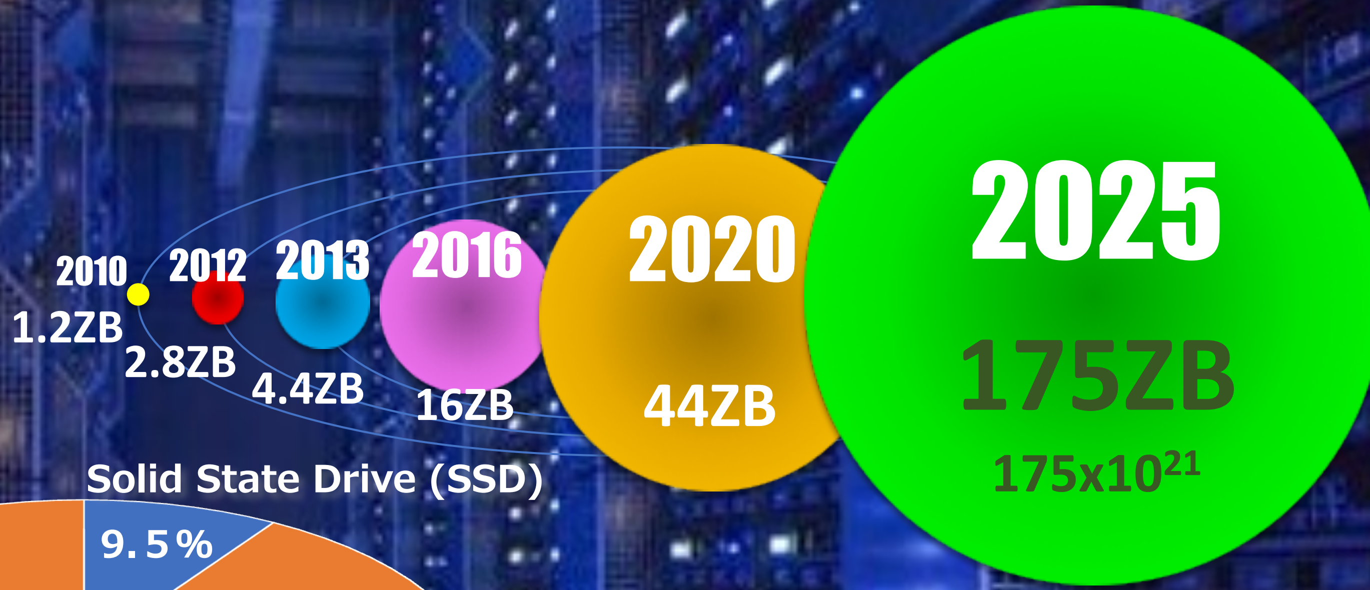
1. Innovation in Magnetic Storage Technology, Perpendicular Magnetic Recording (PMR)
2. Data in Brain Neuroscience, Features, and Issues
3. Close Unification of Data and Computing
4. 3-D Visualization of Neural Structure
5. Summary

Power of Perpendicular Magnetic Recording

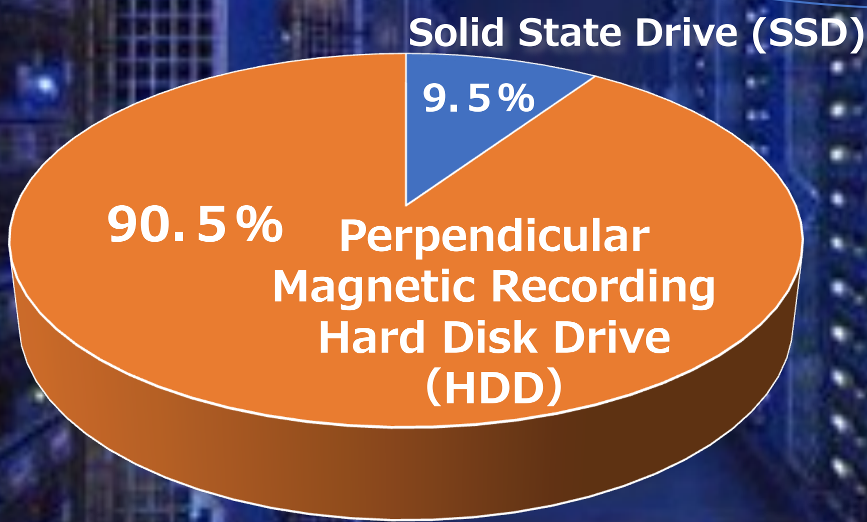
- Personal media player
  - ✓ Larger than 3,500 PB of tiny HDD are playing music
  - ✓ Creates contents business through Cloud



Data Explosion in the World



Source: David Reinsel, John Gantz, John Rydning, Data Age 2025 The Digitization of the World From Edge to Core, An IDC White Paper - #US44413318, 2018



Total Shipment Capacity of HDD + SSD : 5.23ZB (2015~2020)

Source: Based on IDEMA Japan Seminar Presentation by Techno System Research (October, 2020)

Observation of Perpendicular Magnetization

- Perpendicular magnetization was observed in longitudinal recording media.

Circular Magnetization Mode

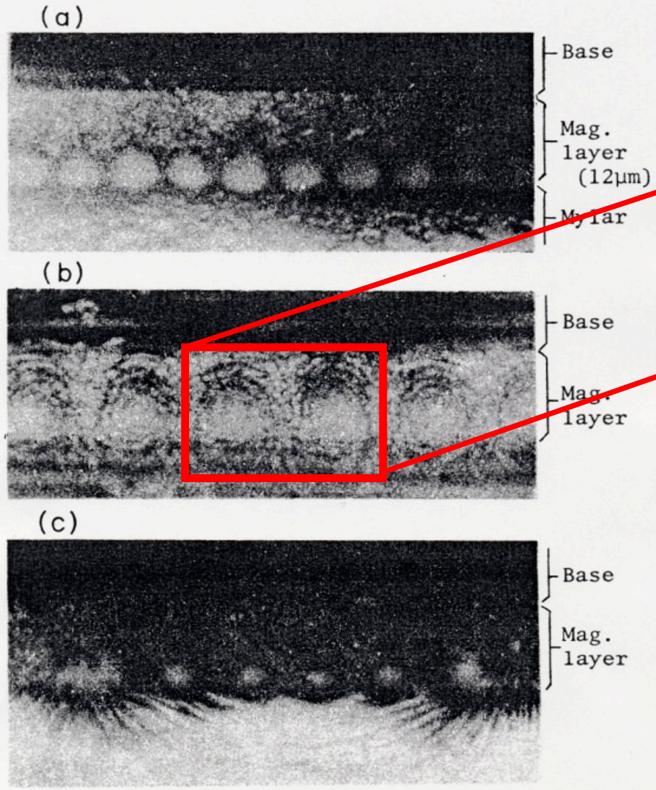
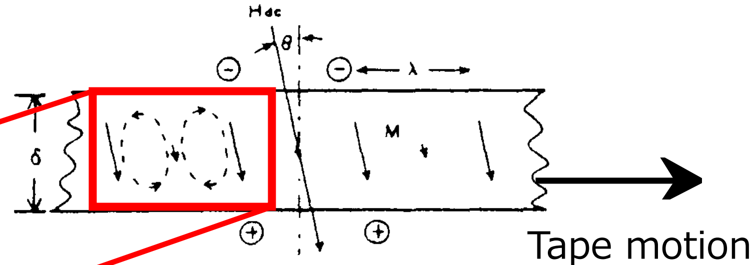


Fig.1 Bitter patterns of (a) no bias, (b) AC bias and (c) NRZ recording (12 bits).

S. Iwasaki and K. Takemura, "An analysis for the circular mode of magnetization in short wavelength recording," IEEE Trans. Magn., vol. MAG-11, no. 5, pp. 1173-1175, Sep. 1975

Method of mode transformation



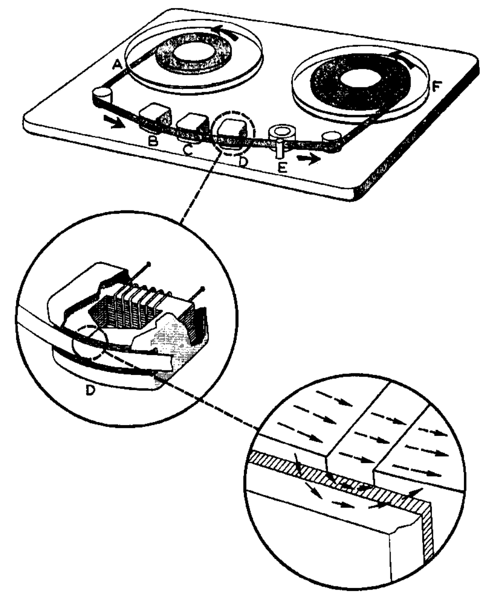
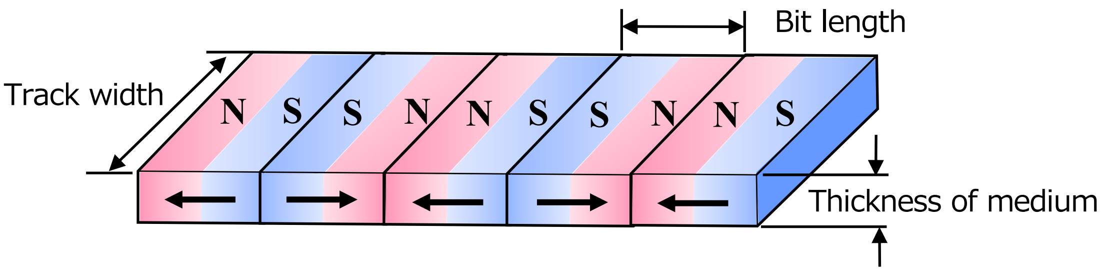


Fig. 1.1. Representation of basic tape recording and reproducing system.

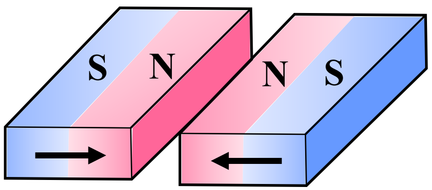
Longitudinal and Perpendicular Recording

- Demagnetization field "enhances" the perpendicular magnetization in high recording density.
- Thick recording layer stabilizes the magnetization, which suppresses thermal demagnetization.

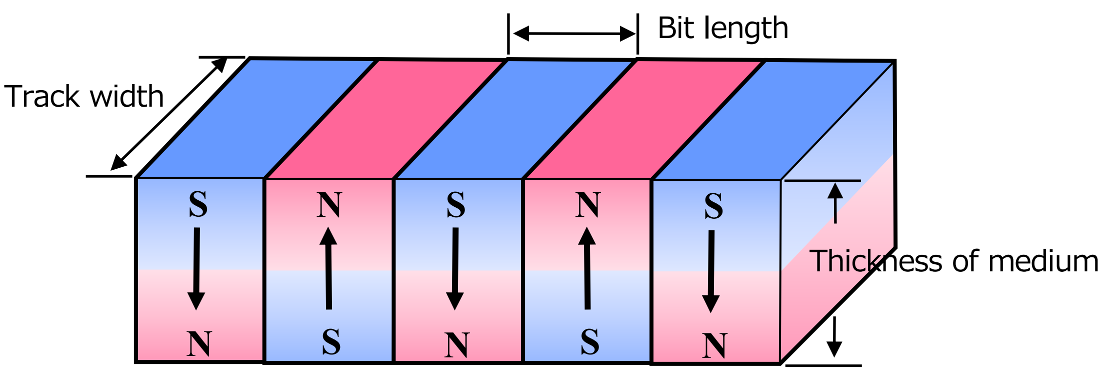
(a) Longitudinal (in-plane) magnetic recording



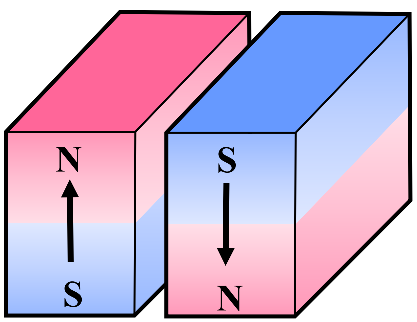
Unstable in high density



(b) Perpendicular magnetic recording



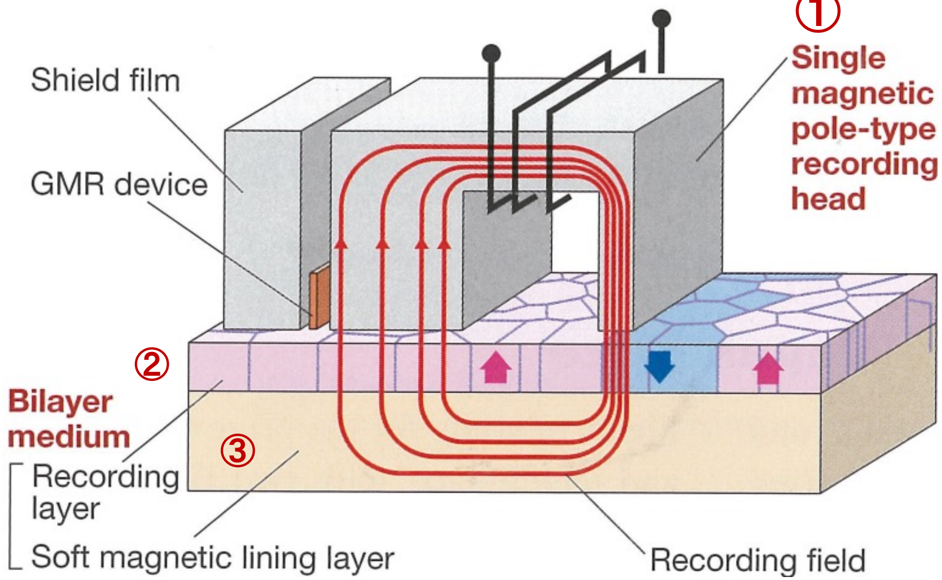
Stable in high density



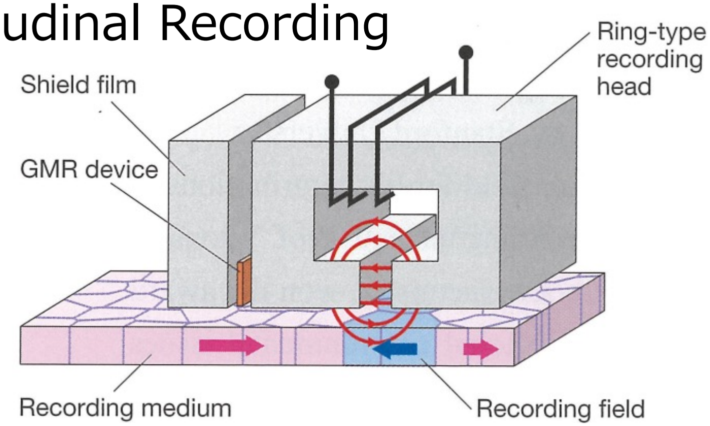
S. Iwasaki, Y. Nakamura, IEEE Trans. Magn., MAG-15, 1272, 1980

Invention of Perpendicular Magnetic Recording

Perpendicular Recording



Longitudinal Recording

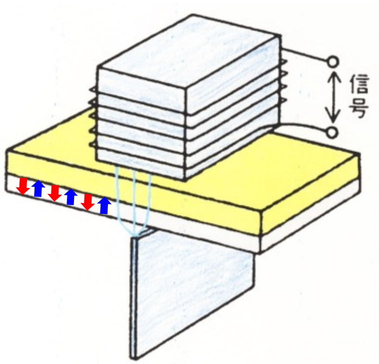


The inventor of perpendicular magnetic recording, Dr. Shunichi Iwasaki (Professor Emeritus, Tohoku University)

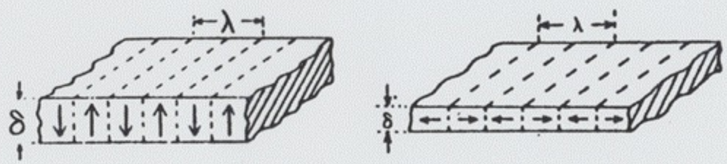
Co-researcher Dr. Yoshihisa Nakamura (Professor Emeritus, Tohoku University)

Complementarity Relationship as Guiding Principle

- Complete picture of magnetic recording by Dr. S. Iwasaki
  - Leads us to high density recording direction

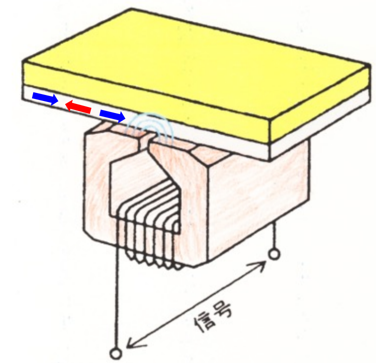


Complementarity relationship between perpendicular and longitudinal magnetic recording.



	a) Perp. Mode	b) Longi. Mode
	$\lambda \to 0 \quad H_d \to 0$	$\lambda \to 0 \quad H_d \to 4\pi M$
Head	Single pole-type	Dipole (ring)-type
Medium	Perp. Anisotropy Thick δ High Ms, High Hc	Longi. Anisotropy Thin δ Low Ms, High Hc
Signal	Digital (Sat.)	Analog (non-Sat.)
Rec. Method	Modulation (FM, PCM)	AC Bias Method
Erase	DC Field	AC Field

S. Iwasaki, IEEE Trans. Magn., vol. MAG-15, 71 (Jan. 1980).

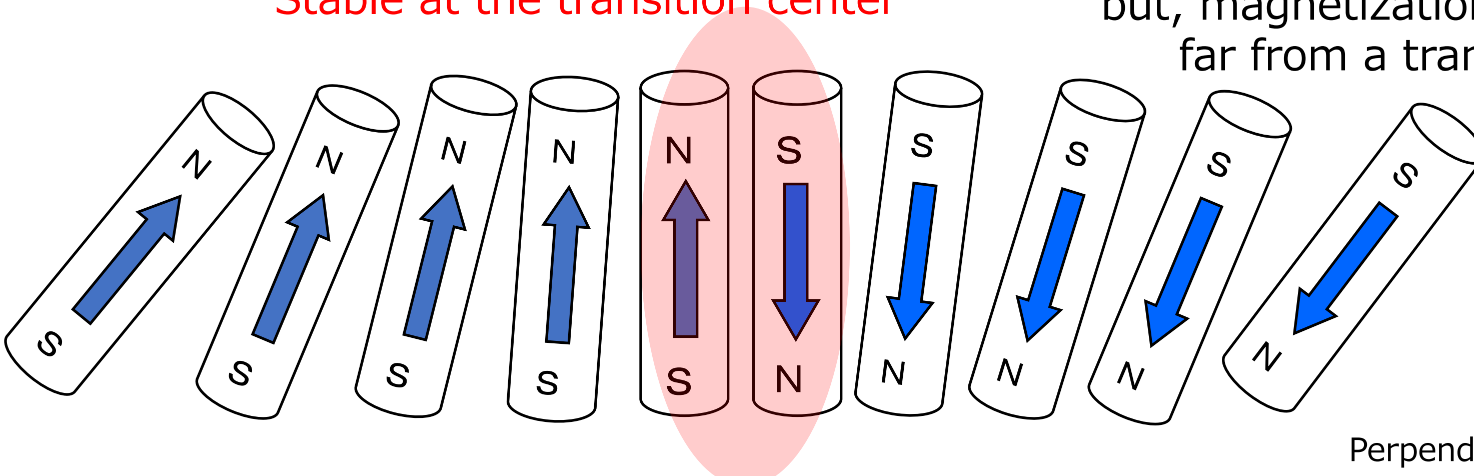


A Big Issue of Magnetization

Originally proposed CoCr PMR medium

Stable at the transition center

but, magnetization declines far from a transition

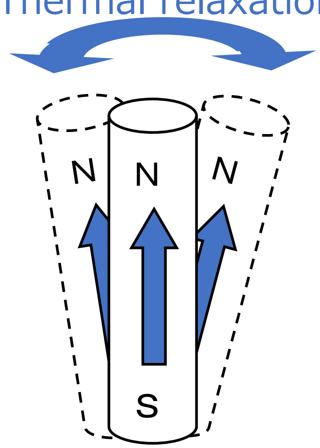


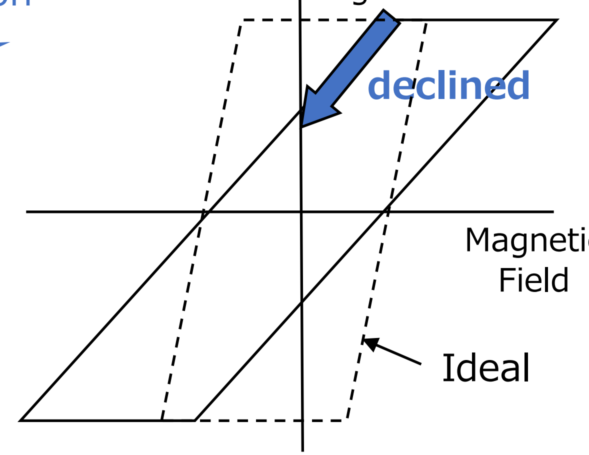
① Magnetic Anisotropy is not large enough

② Magnetization declines due to self-demagnetization

= MH loop low squareness

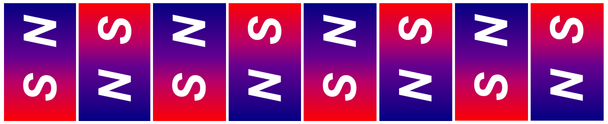
week to thermal relaxation





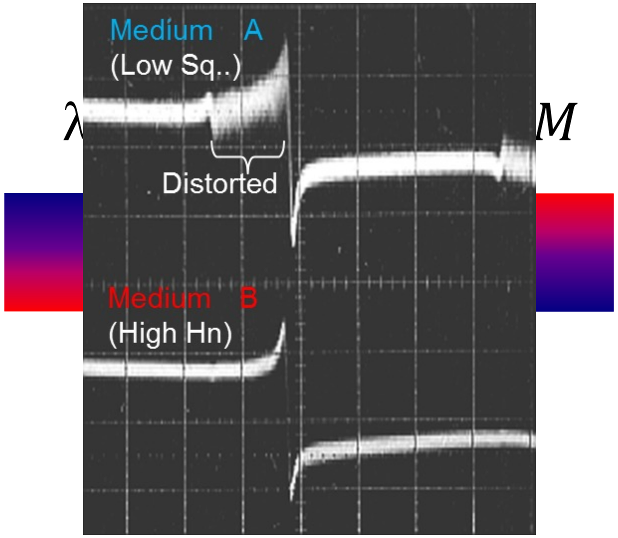
Bird's Eye View

	Perpendicular	**Longitudinal**
Nature*	$\lambda \to 0 \quad H_d \to 0$	$\lambda \to 0 \quad H_d \to 4\pi M$

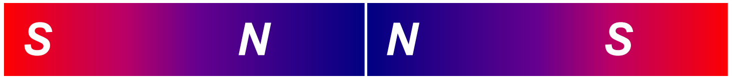


Worst Case

Signal from an Isolated Transition



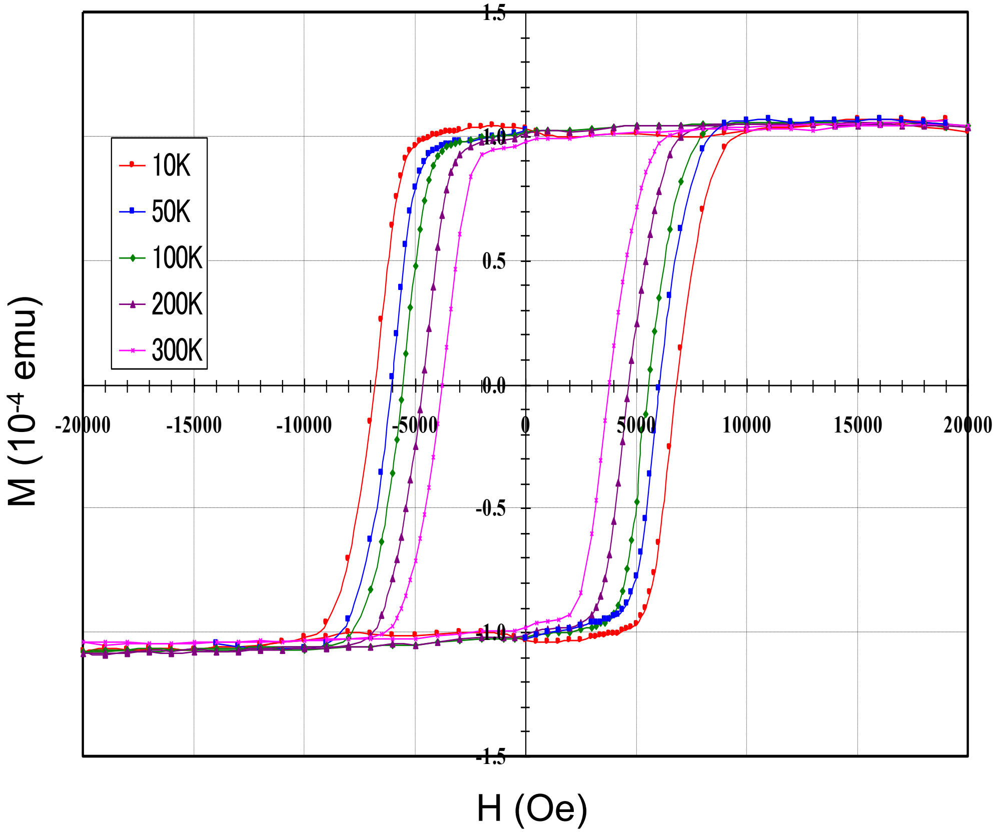
$\lambda \to \infty \quad H_d \to 0$

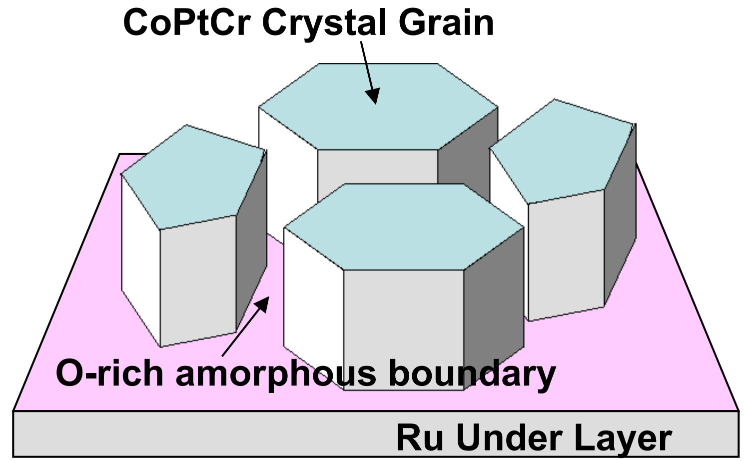


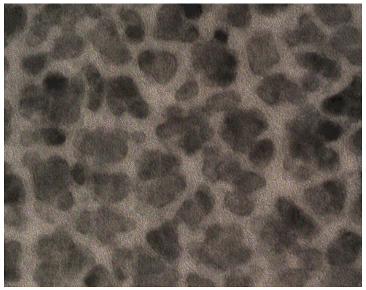
A. Takeo, S. Oikawa, T. Hikosaka, Y. Tanaka, "A new design to suppress recording demagnetization for perpendicular recording", IEEE Trans. Magn., vol. 36, no. 5, pp. 2378-2380, Sep. 2000

High Magnetic Energy Perpendicular Media

- Developed CoPtCrO medium on Ru underlayer
- High squareness by exchange coupling





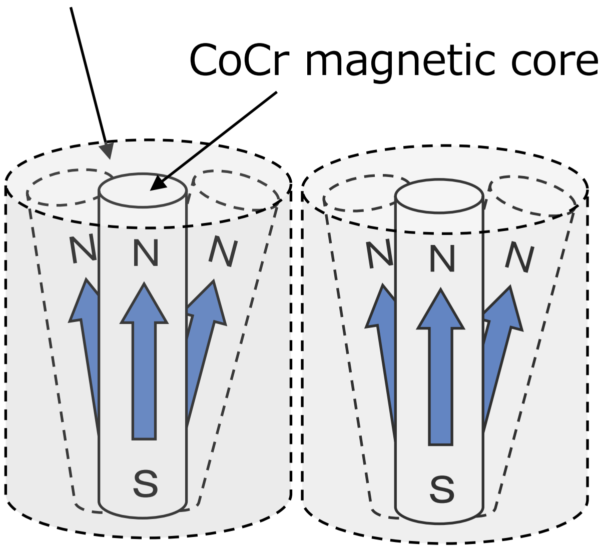


Fine granular microstructure (TEM Image)

Y. Tanaka, T. Hikosaka, Perpendicular recording with high squareness CoPtCrO media, J. Magn. Magn. Mater., 235, pp.253-258, 2001

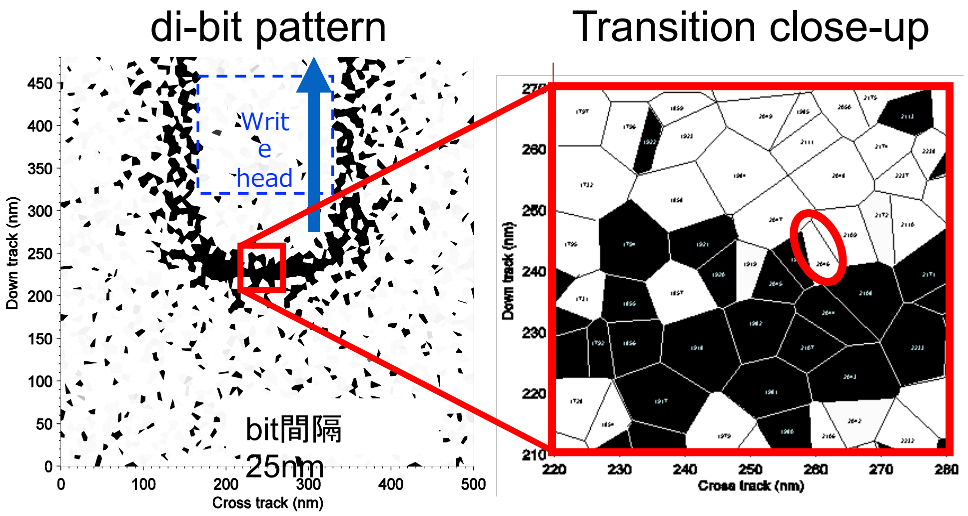
Unstable Magnetic Structure

Traditional CoCr Medium



Cr reduces magnetic anisotropy of the magnetic core resulting in unstable magnetization.

Magnetization reversal behavior of a single particle in writing process

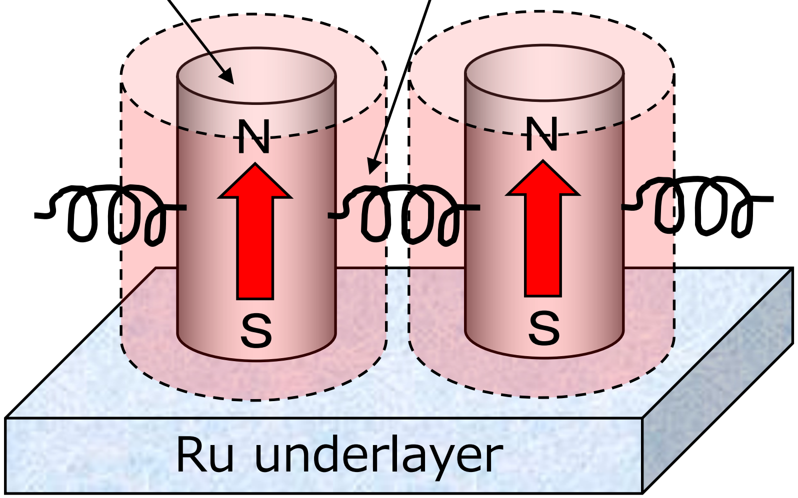


Stable PMR Structure

New CoPtCr/Ru PMR medium

High anisotropy energy CoPtCr magnetic core

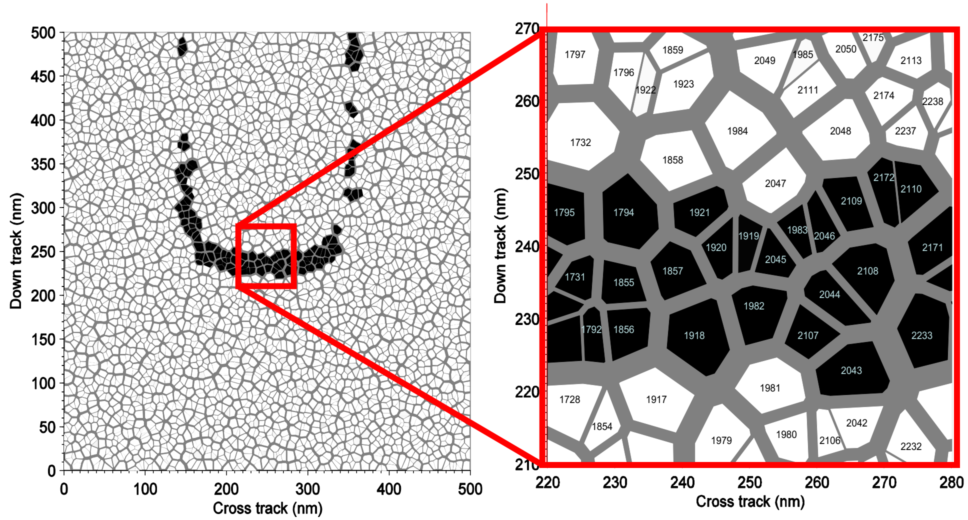
Oxide-rich grain boundary with exchange coupling



High H_n medium forms stable transitions

di-bit pattern

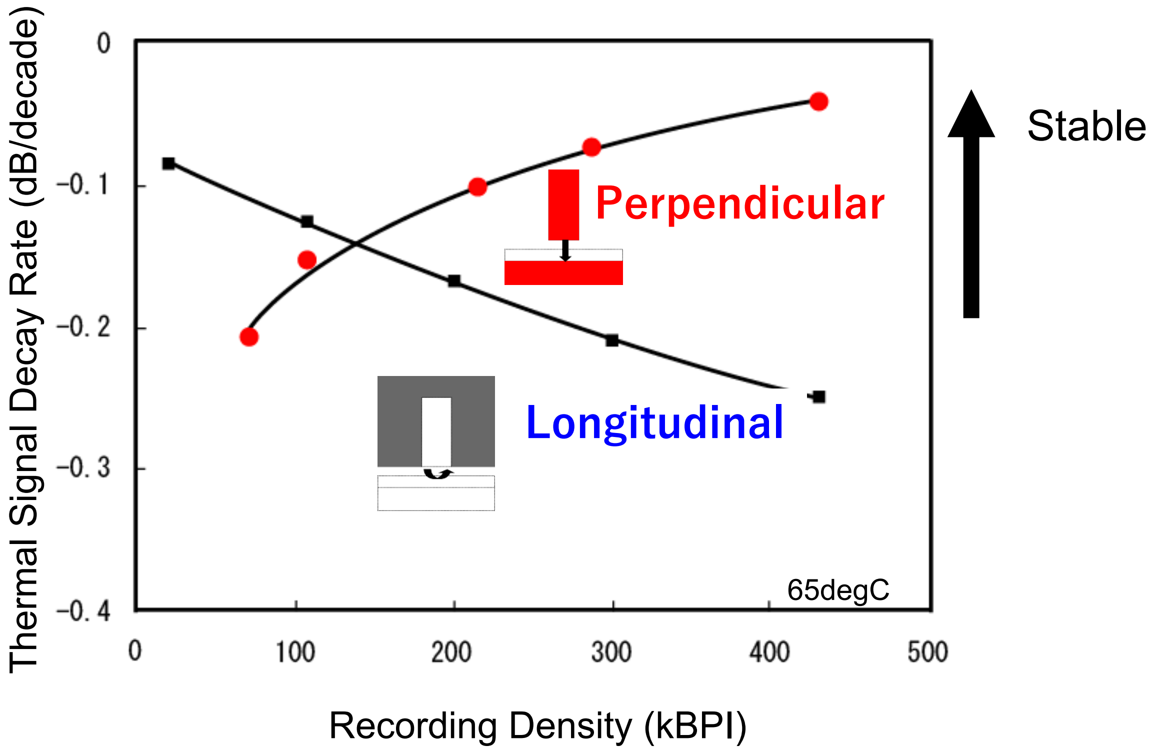
Transition close-up



- Grains supports each other to make stable structure
- Oxygen, we have avoided, plays a big role to isolate CoPtCr magnetic grains

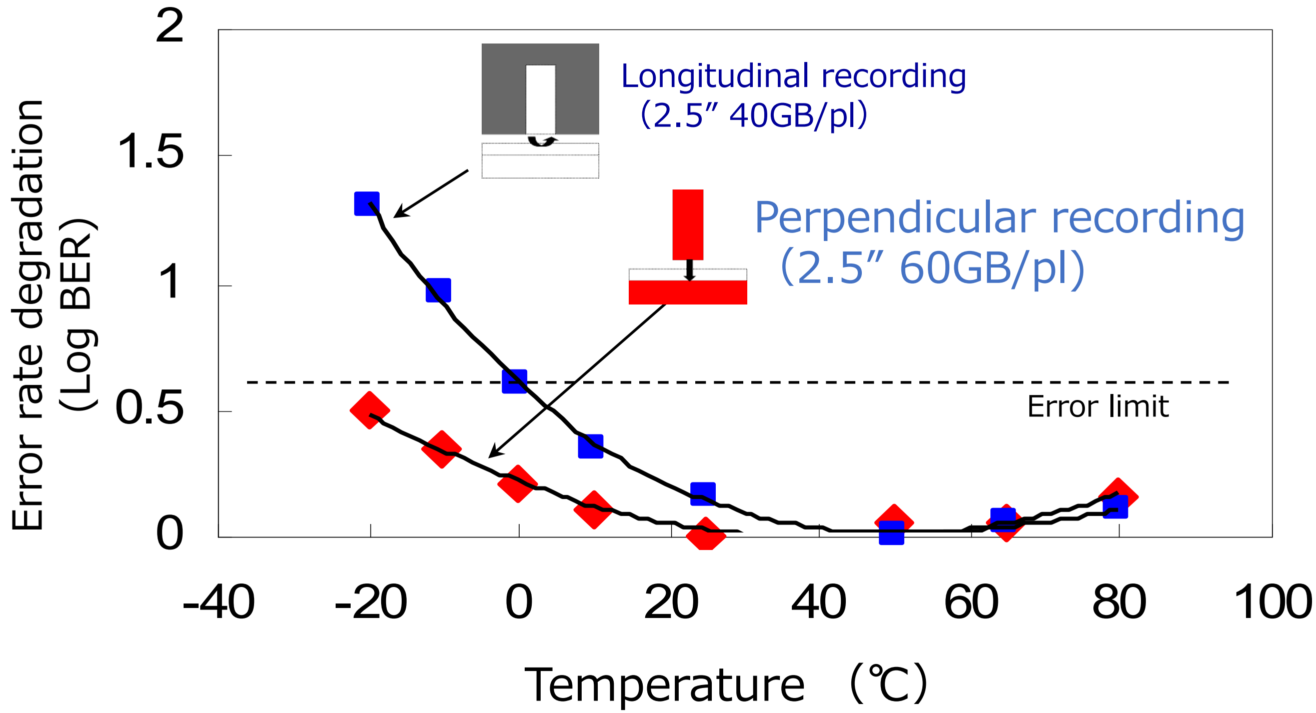
Complementarity in Thermal Stability

- Perpendicular: Higher density, better thermal stability
- Intrinsically Opposite Characteristics



Superior Write Performance at Low Temp

- PMR is very robust even at low temperature
- PMR can expand the temperature range by 20 degrees
- Medium in write magnetic flux path



Complementarity as Backbone Principle

Original findings by Dr. S. Iwasaki

	Perpendicular	**Longitudinal**
	$\lambda \to 0,\ H_d \to 0$	$\lambda \to 0,\ H_d \to 4\pi M$
Head	Single pole-type	Dipole (Ring)-type
Medium	Perpendicular anisotropy	Longitudinal anisotropy
	Thick d	Thin d
	High M_s, High H_c	Low M_s, High H_c
Signal	Digital (saturation)	Analog (non-saturation)
Rec. Method	(FM, PCM)	AC bias method
Erase	DC field	AC field

Performance-related relationship in HDD integration by Y. Tanaka*

Media	High squareness	Low squareness
	With soft underlayer	Recording layer only
Thermal Stability	Good at high density	Good at low density
Write Process	Medium in write flux path	Medium outside of path
	Low spacing sensitivity	High spacing sensitivity
	Sharp transition	Broad transition
Read Process	High output	Low output
	Narrow reading	Wide reading
Signal	With DC component	Without DC component
Signal Processing Channel	Positive coefficient PRML	Negative coefficient PRML

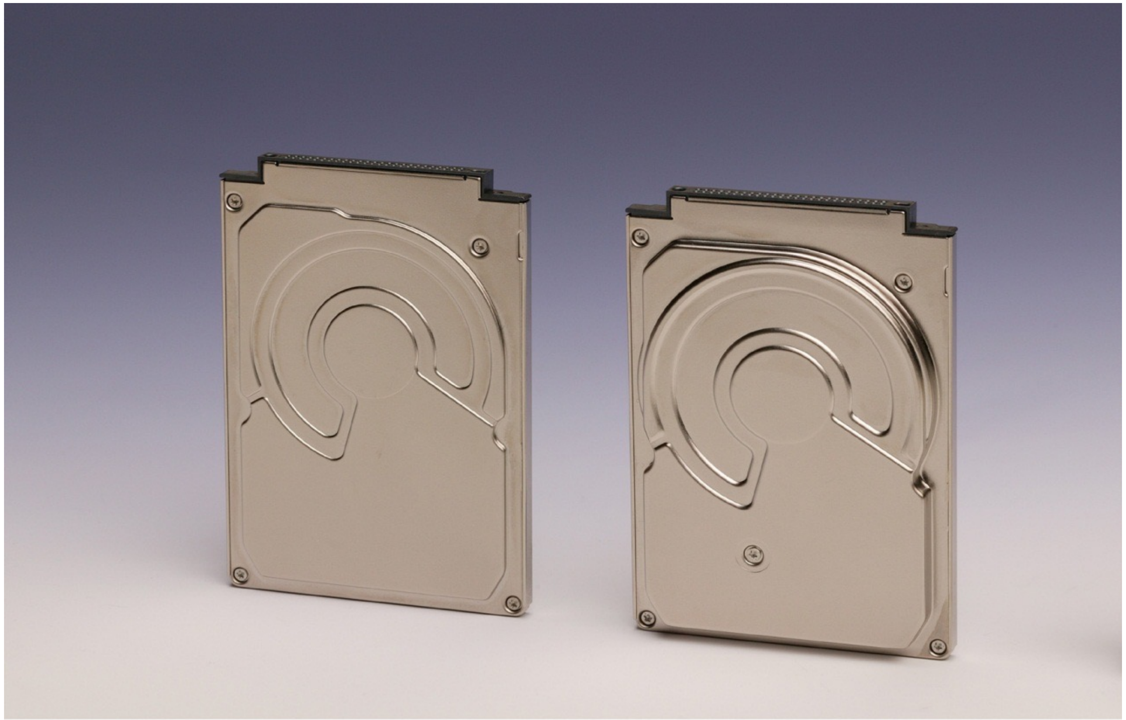
* Y. Tanaka, "Fundamental features of perpendicular magnetic recording and design consideration for future portable HDD integration", IEEE Trans. Magn., vol. 41, no. 10, pp. 2834-2838, Oct. 2005

All Green for Takeoff; All Issues Solved

- Head pole/yoke-related erasure
- External magnetic field sensitivity
- Spacing sensitivity
- Media SNR
- Soft under layer noise
- Thermal relaxation at low BPI
- Too-narrow side erase band

Y. Tanaka, "Fundamental features of perpendicular magnetic recording and design consideration for future portable HDD integration", IEEE Trans. Magn., vol. 41, no. 10, pp. 2834-2838, Oct. 2005

The World First Product of Perpendicular Recording HDD



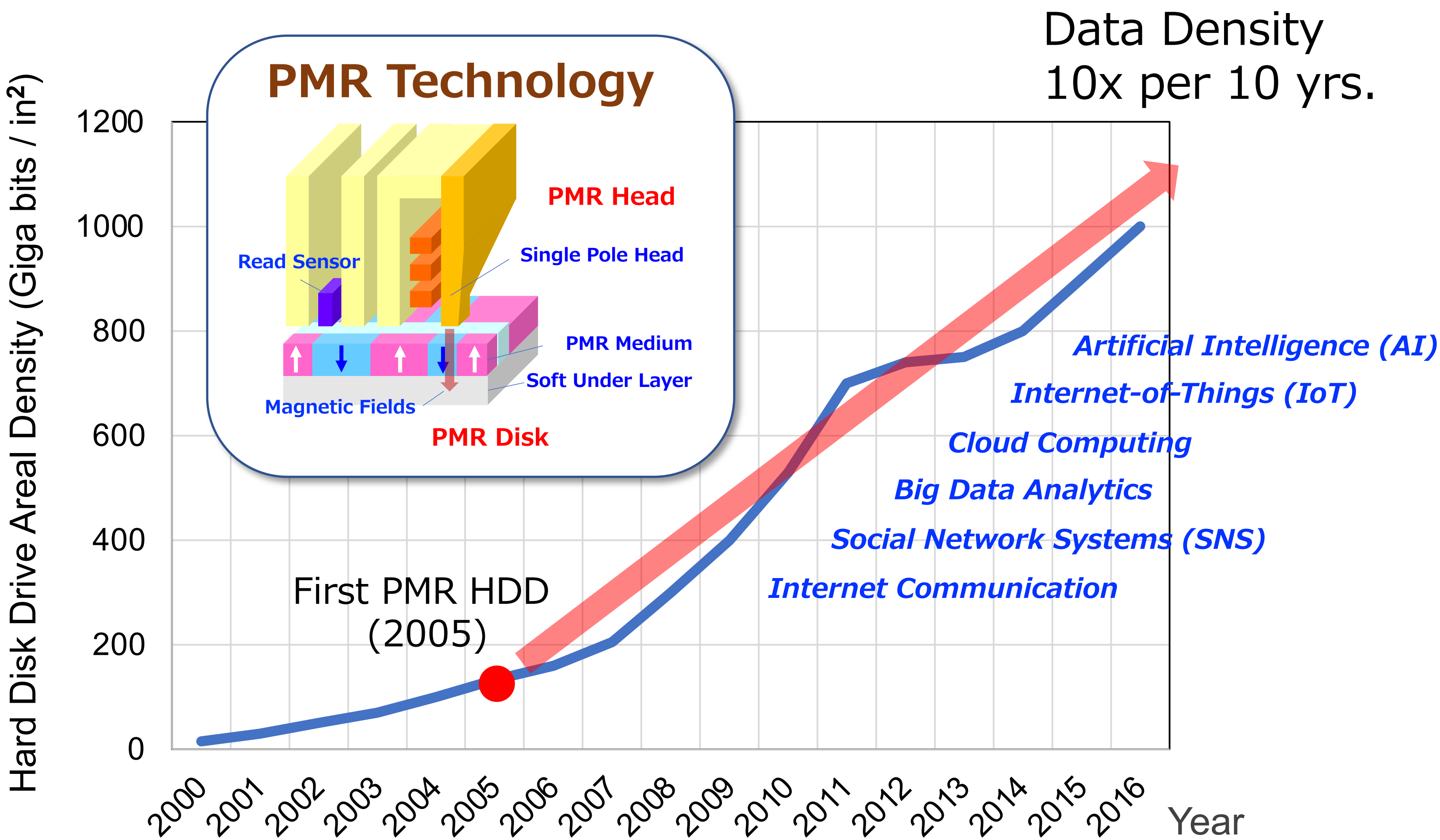
TOSHIBA

1.8”HDD 40GB “MK4007GAL” and 80GB “MK8007GAH” (2005)

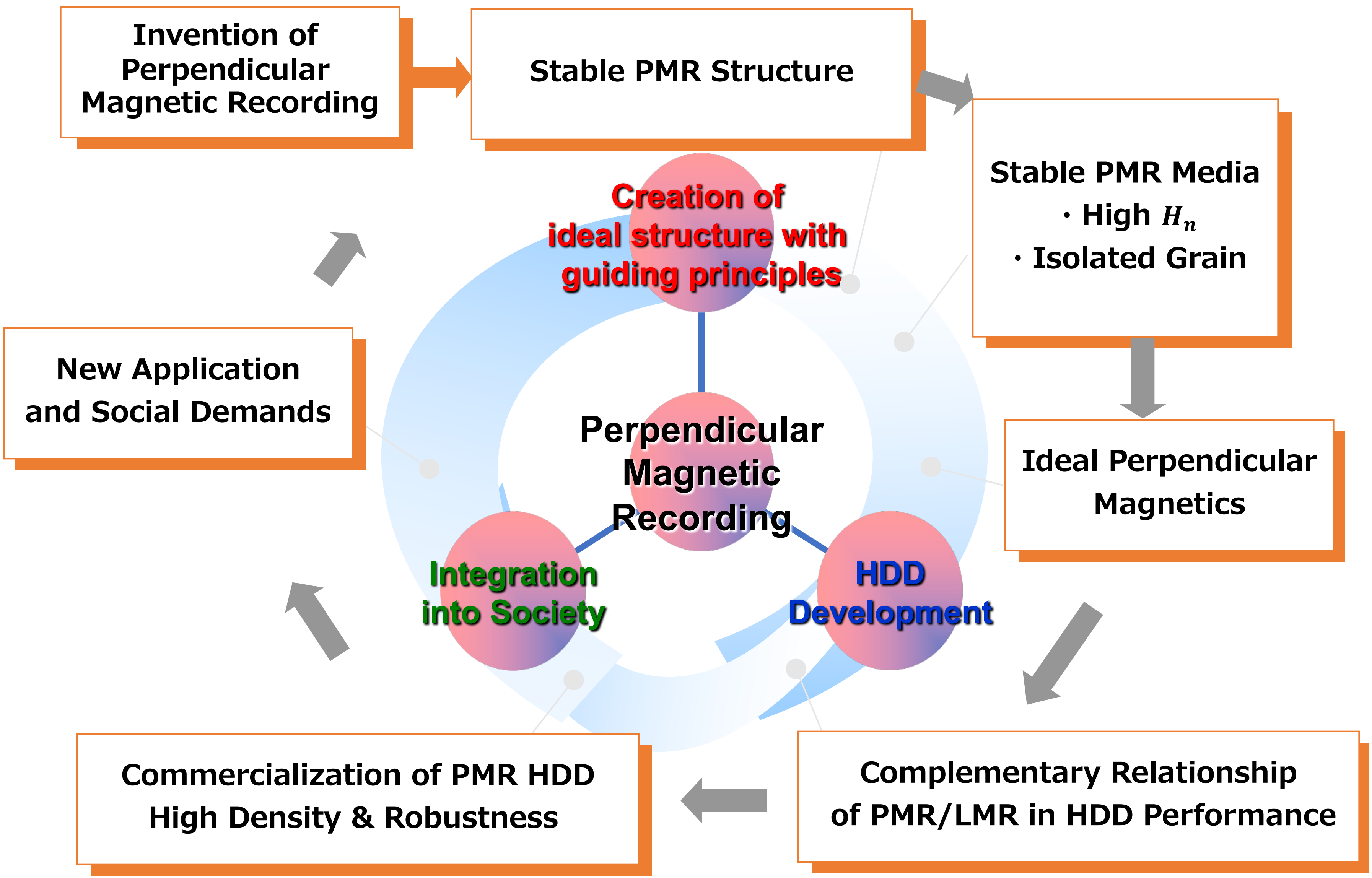
The highest areal density at 133 Giga bits/in^2 at launching

http://www.toshiba.co.jp/about/press/2004_12/pr1401.htm

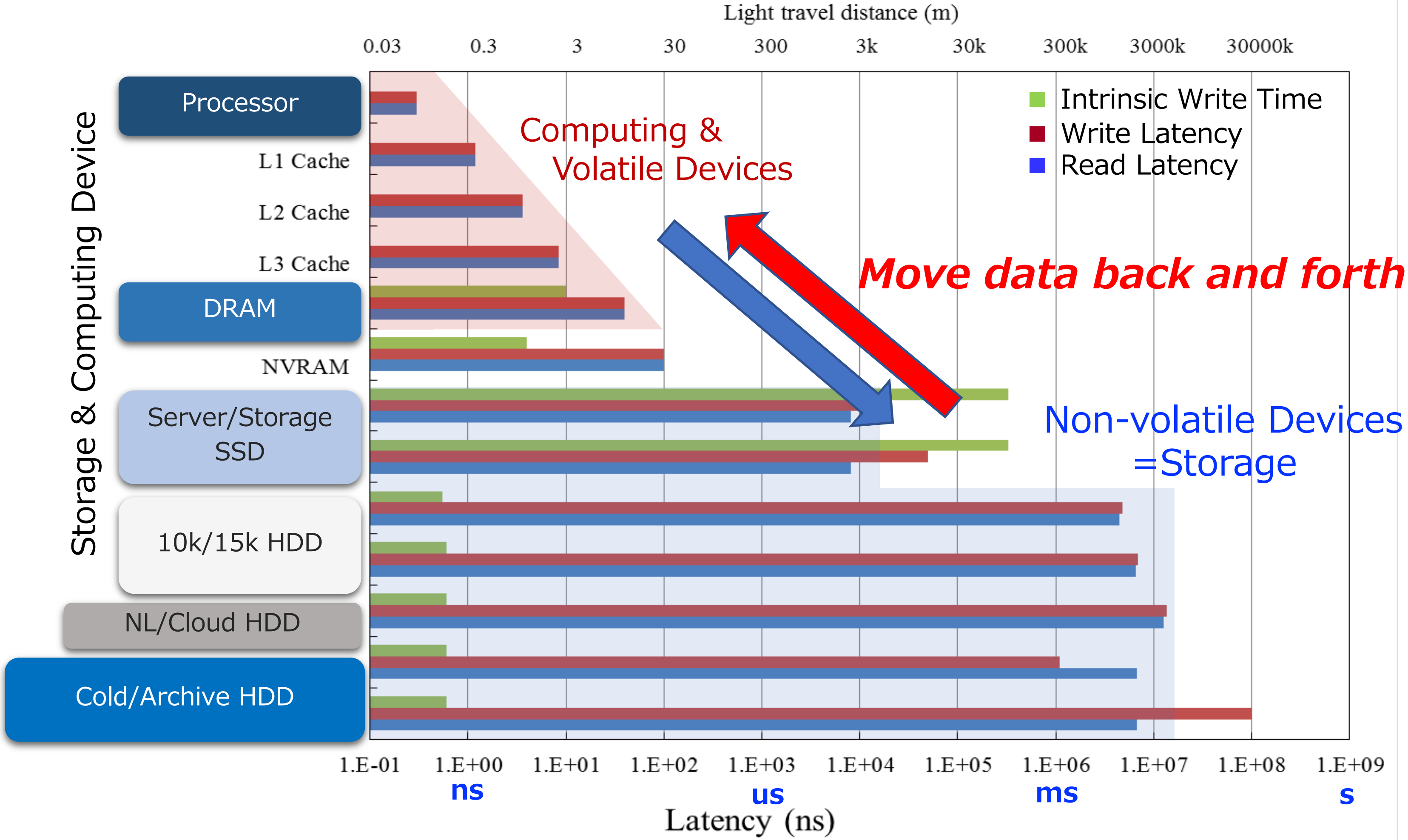
Bigdata Platform by Large Capacity Storage



Innovation Clock of PMR



Look back Current Data Tiers



Yoichiro Tanaka, Characterizing Advanced Recording Technology Assets with Hyper-Scale Applications, IEEE Trans. Magn., Vol.52, No.2, pp.1-4, 2016

Latency Kitchen Model

Convert the latency to light travel distance

Volatile Devices

Processor	**L1 Cache**	**L3 Cache**	**DRAM**
	30cm	**3m**	**10m**

Non-volatile Devices

HDD

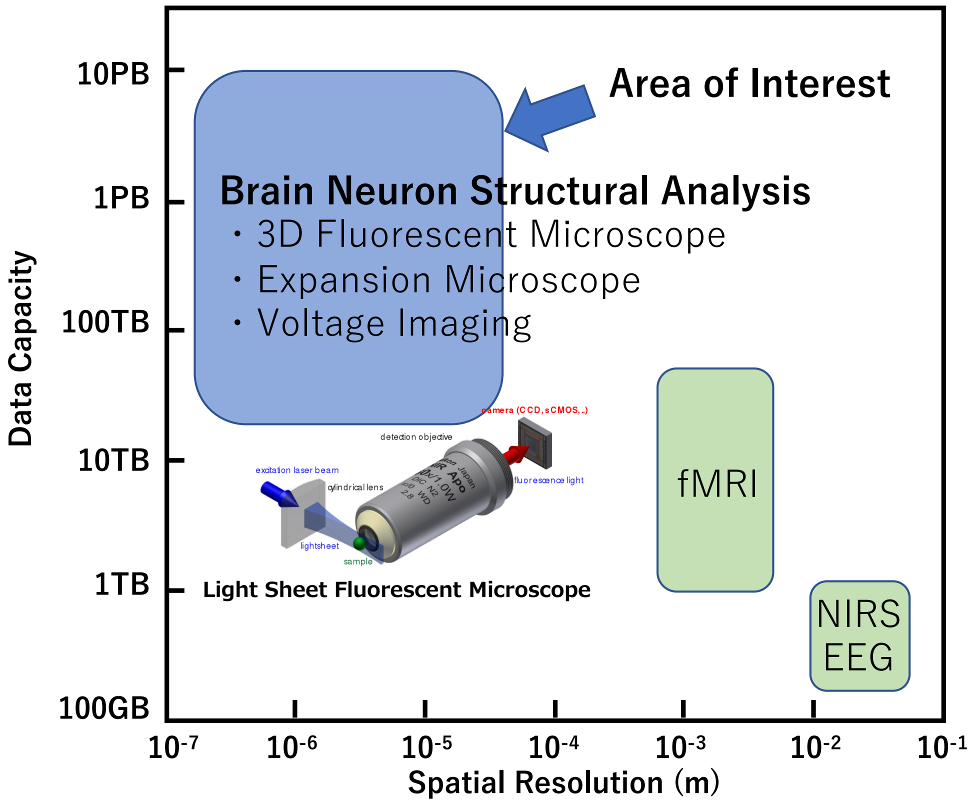
2000km

SSD

3km

Move foods back and forth !

Increased data capacity in microscopic neuron observation

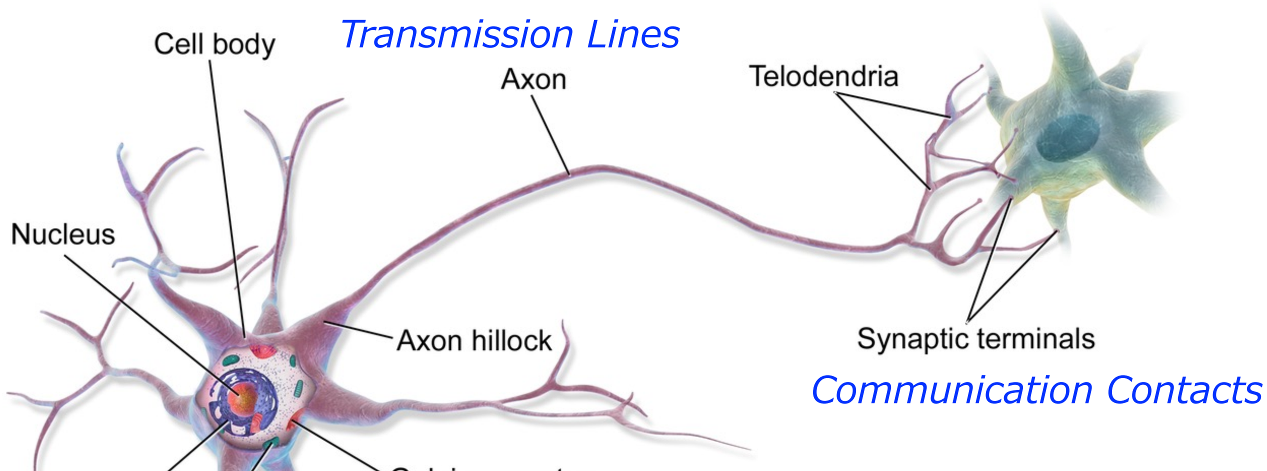


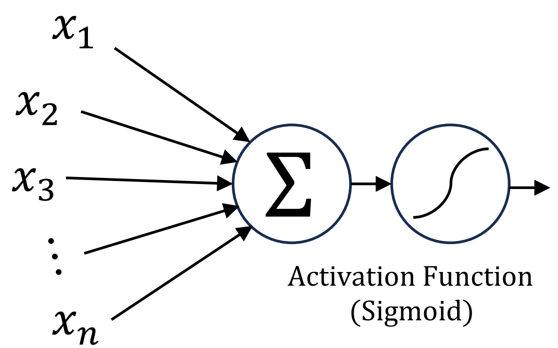
- Long time-sequence data
- Multi-channel data streams from sensors
- Real time data acquisition
- Complex data analytics and visualization
- Highly frequent data access from multiple clients
- Secure store of precious data

*Source : Patric Hagmann, "Chairman's introduction: definitions and basic imaging technique", NH7: The human connectome: a comprehensive map of brain connections, ECR 2014 (Courtesy of Dr. Hitoshi Yamagata, Currently Canon Medical Co.)

Yoichiro Tanaka
Research Institute of Electrical Communication, Tohoku University

Neurons and Synapses





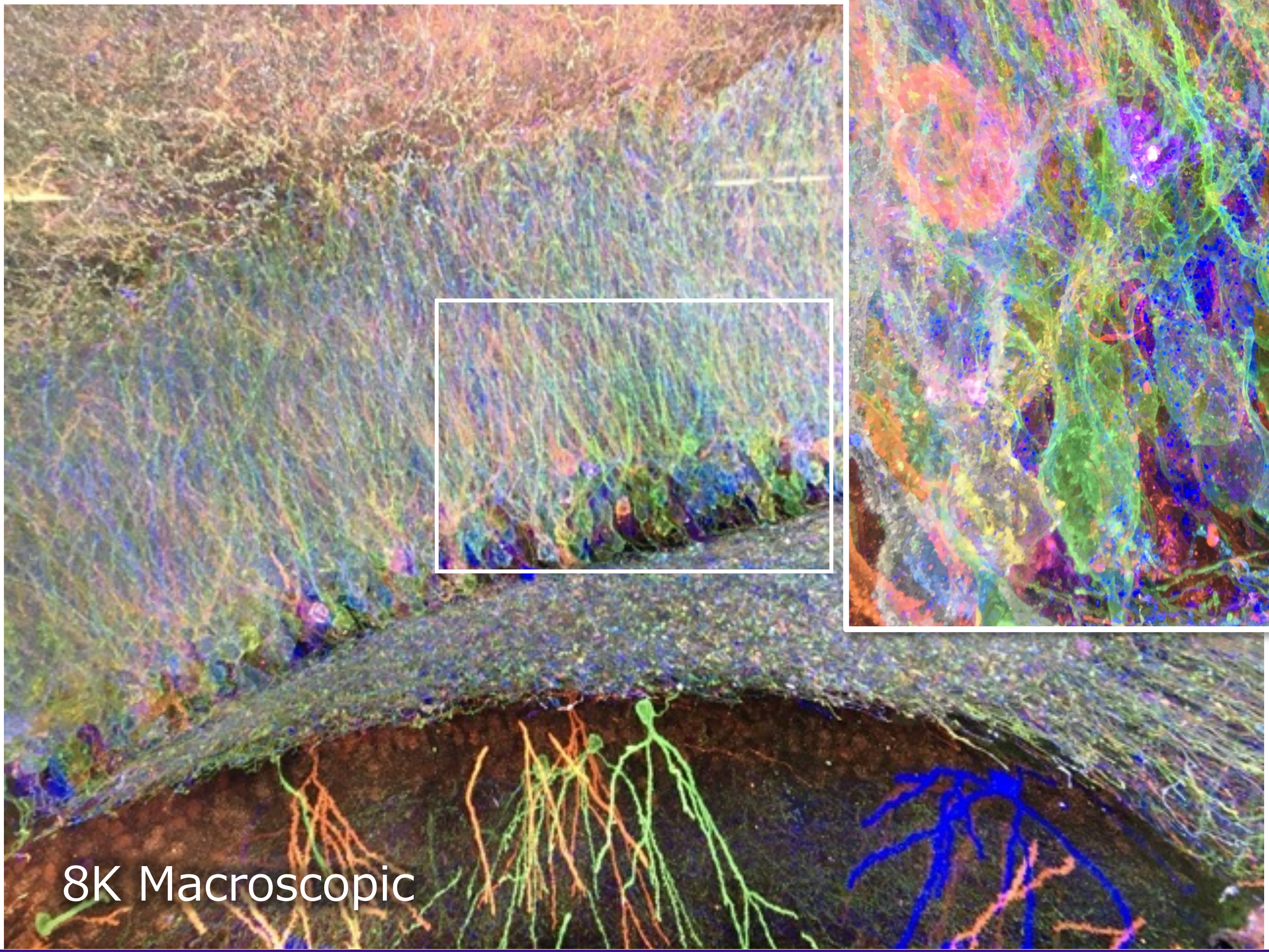
Neurons 100,000,000,000 approx.
Synapses 150,000,000,000,000 approx. in a human brain system

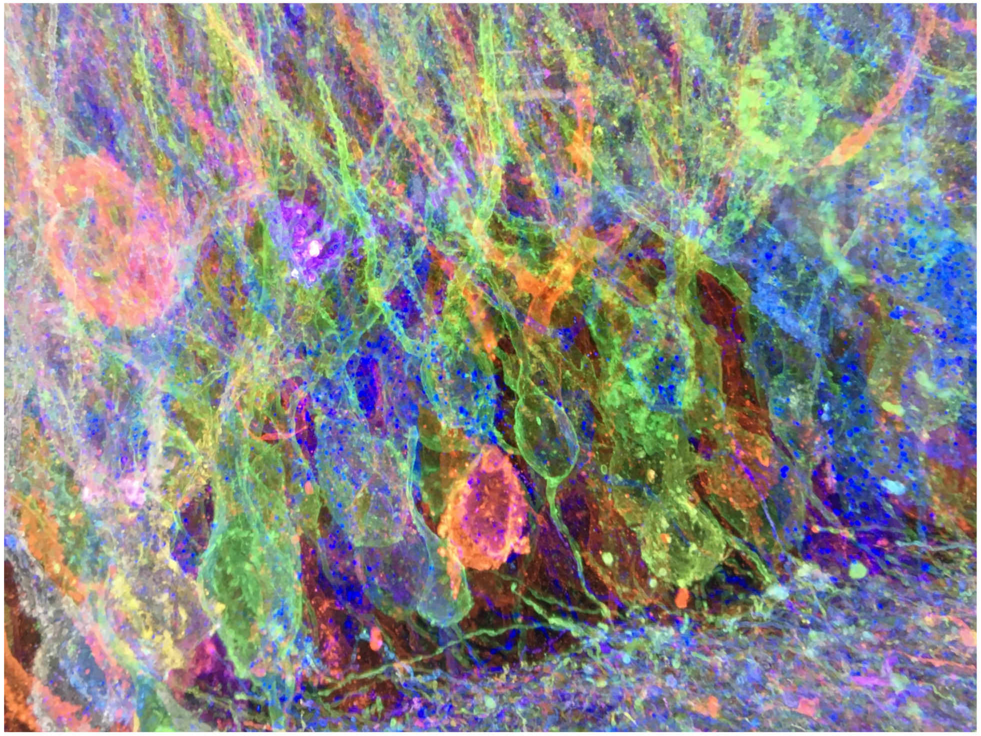
https://en.wikipedia.org/wiki/Neural_circuit

Multi-scale 3D Analytics of Brain Structure

From synaptic contacts (10^{-9}m) to whole brain (10^{-1}m)

Mouse hippocampus neuron structures





Macroscopic & microscopic 8K images (85" display) of mouse hippocampus neuron structures, scanned by Light Sheet Microscope
Data set size; 6TB ~ 60TB
Specimen: 1.5mm x 0.8mm x 0.2mm^t
2000 scans per 0.2mm^t (100nm^t each)
Scan section 80um x 8.3um

(Brain Data Center Project; MIT, NHK, Toshiba Memory, Y. Tanaka, July 2017)

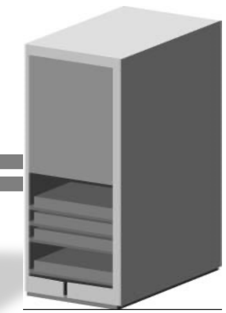
Brain Sensing:
Micro-scale Structural Mapping

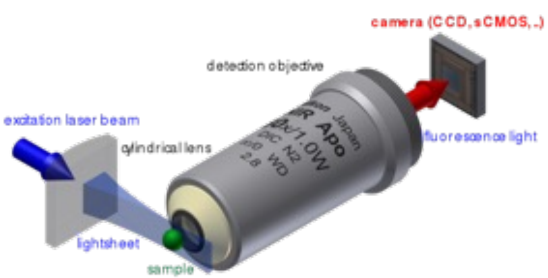
Raw Data Storage → **Data Analysis** → **Visual Mapping**

Fluorescent Microscope

Electrodes Probing

→ **High Speed Record System** → **Scalable Store & Compute**





Light Sheet Fluorescent Microscope

4.5PB / hr (>1PB / session)

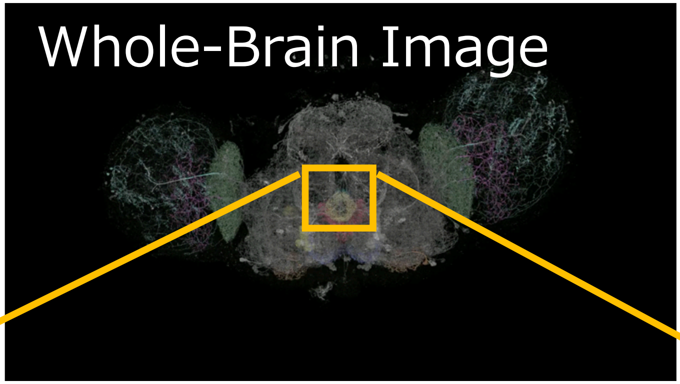
10Tbps = 1G pixels x 1000fps x 10bit

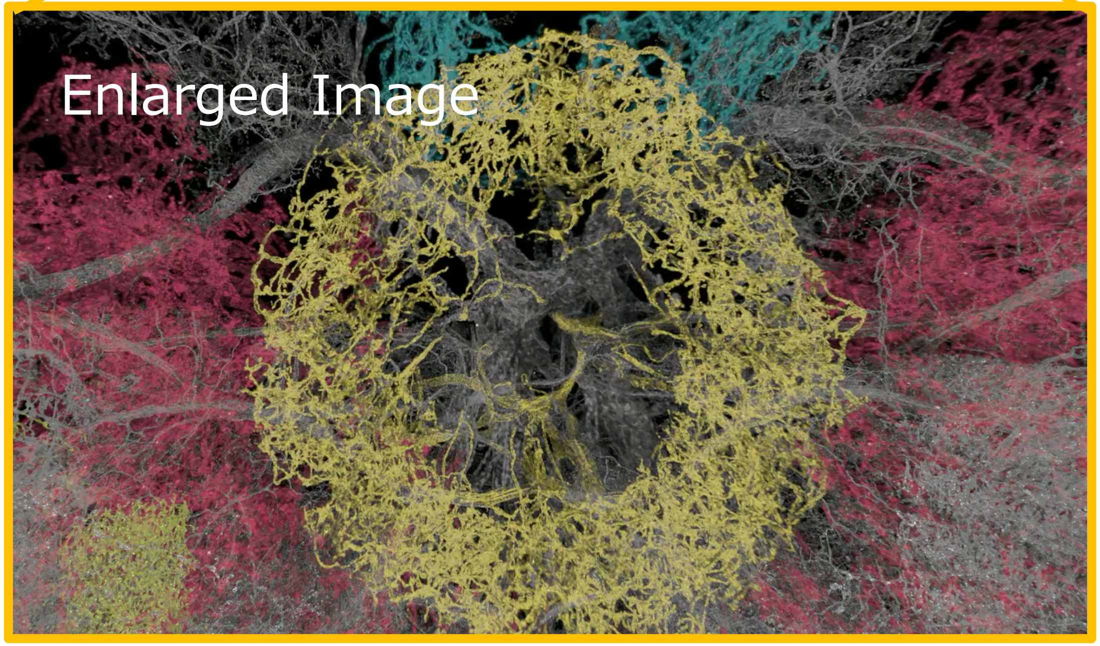
Visualization of Whole Brain

A 3D-image sample from Drosophila (fruitfly) whole-brain neuron structures observed by Ex-LLSM[1).

Only selected neurons which are about 10% of whole 100,000 ones are shown.

The images are retrieved from disaggregated data store using 3D-visualization tool[2) in a compute node.



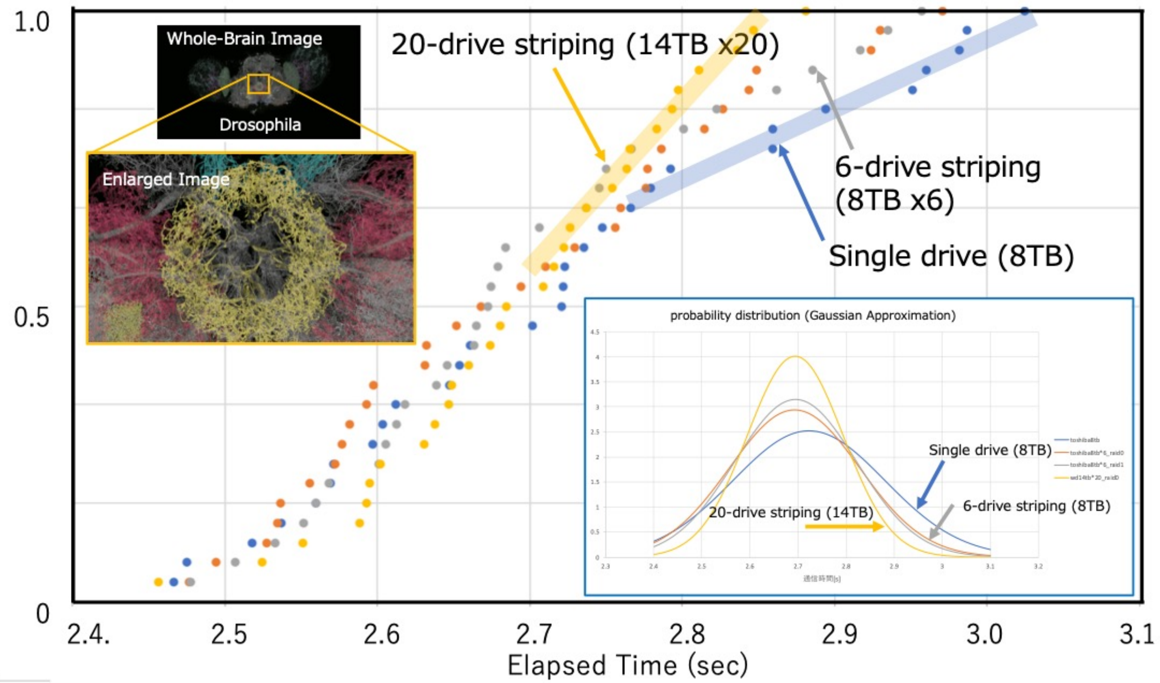


1) Ruixuan Gao, et al., “Cortical column and whole-brain imaging with molecular contrast and nanoscale resolution”, Science 363, 245, 2019
2) Supported by MIT and Kioxia Corp.

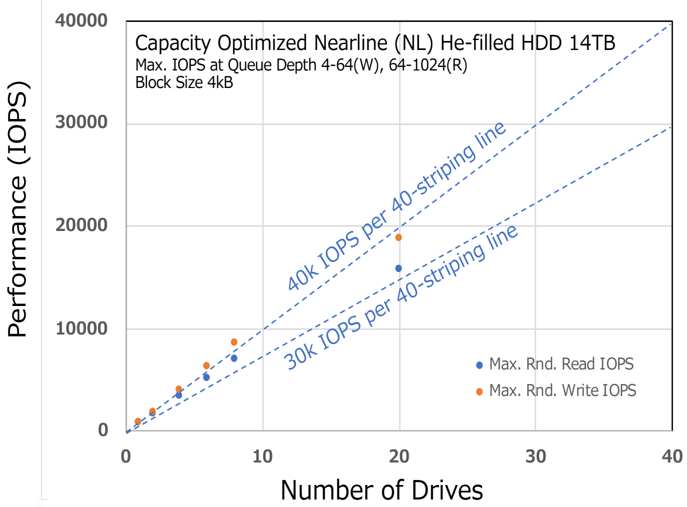
Data Access Performance of 3D Neuron Datasets

Multi-parallel striping in disaggregated storage pool suppressed the tailing of access latency.

Access latency distribution of retrieving Drosophila's whole brain datasets from disaggregated files



HDD 40-striping:
Random Performance 30k~40k IOPS

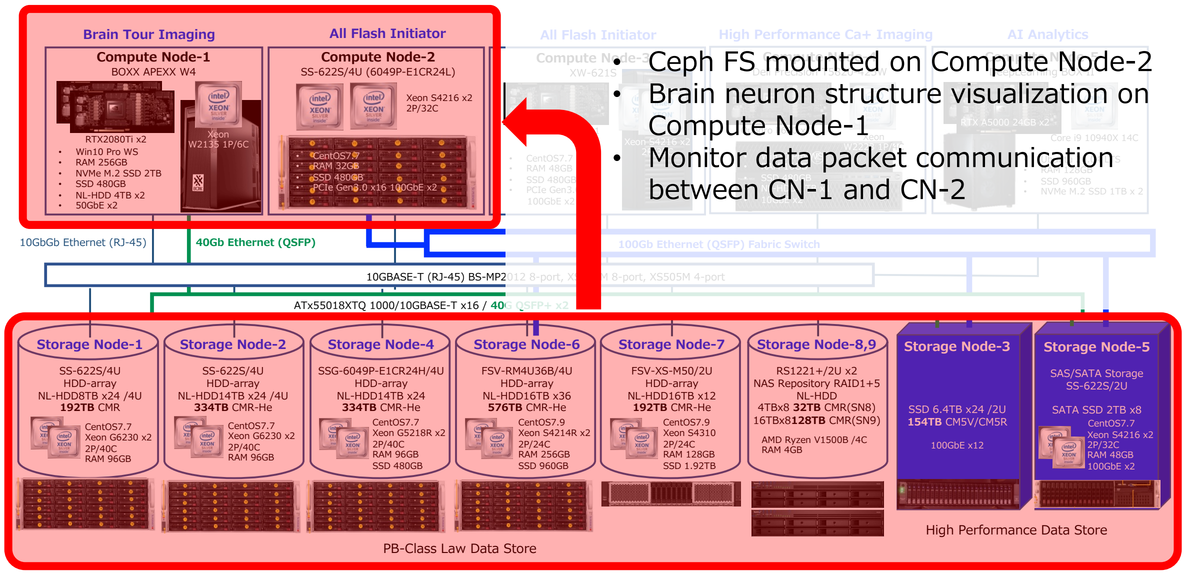


Object Storage System "Ceph"

Compute and Storage Testbed

2PB Storage (HDD,SSD), 6GPU/21CPU, 1TB memory

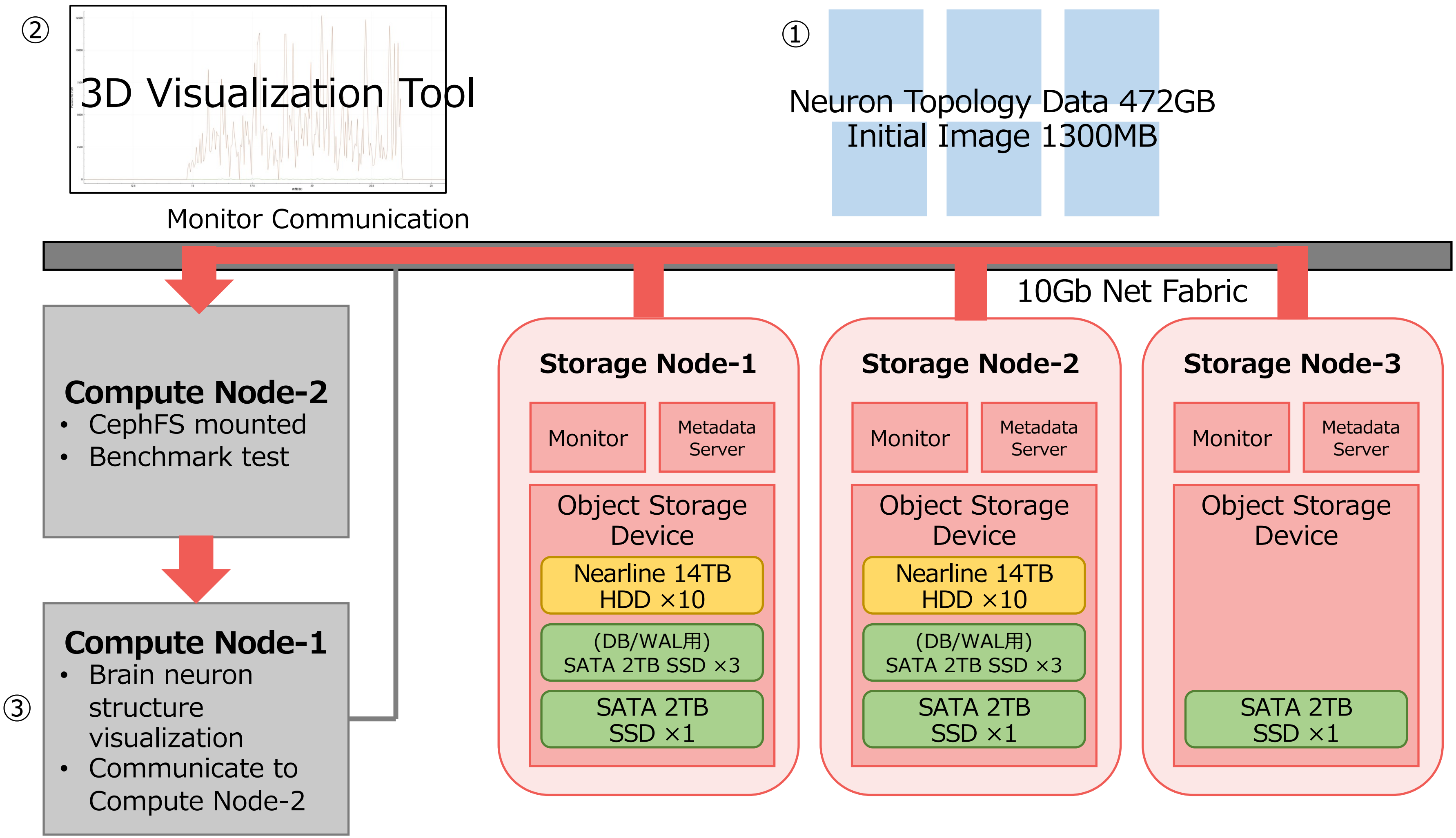
ceph



- Ceph FS mounted on Compute Node-2
- Brain neuron structure visualization on Compute Node-1
- Monitor data packet communication between CN-1 and CN-2

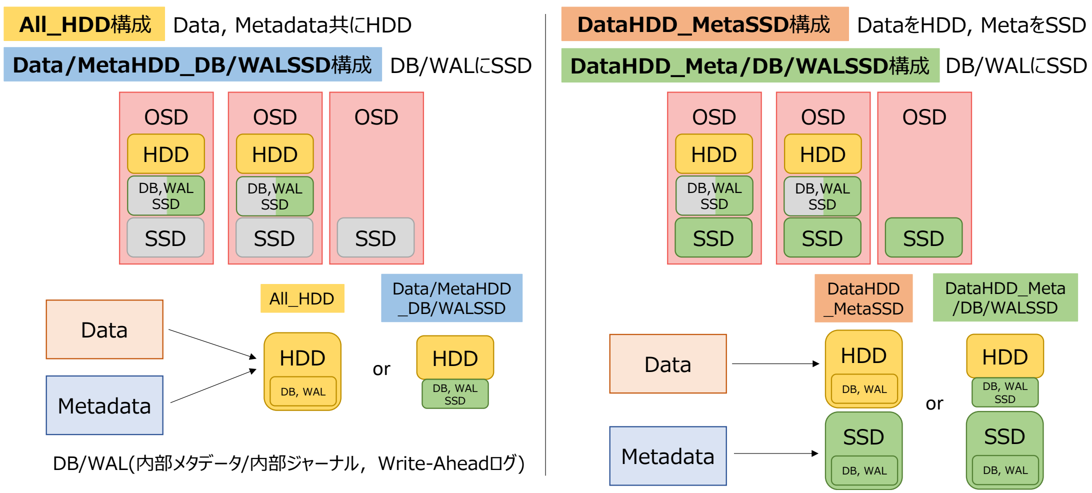
RADOS
Reliable Autonomic Distributed Object Store

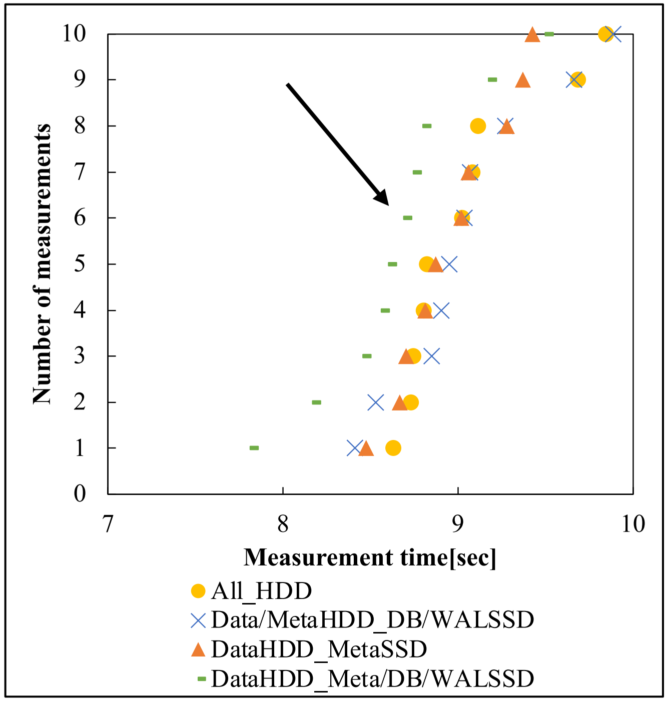
Data Access Performance



Backend data stored in SSD

Performance of data access to HDD improved by storing DB and WAL in SSD

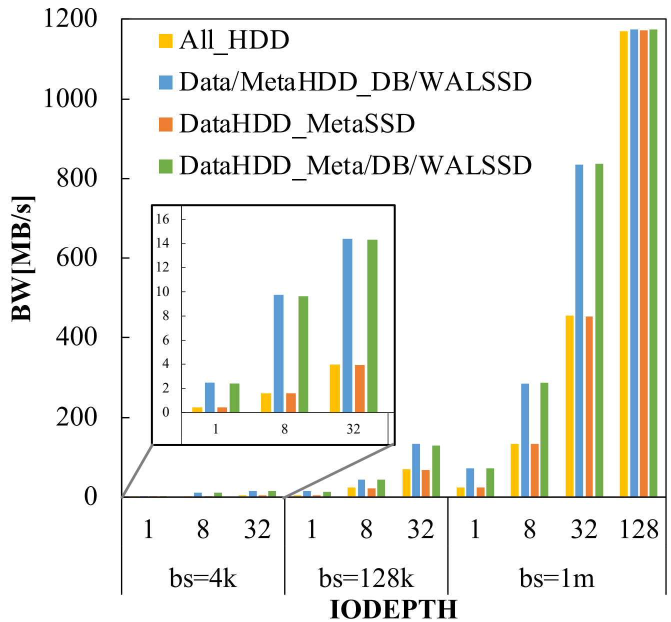




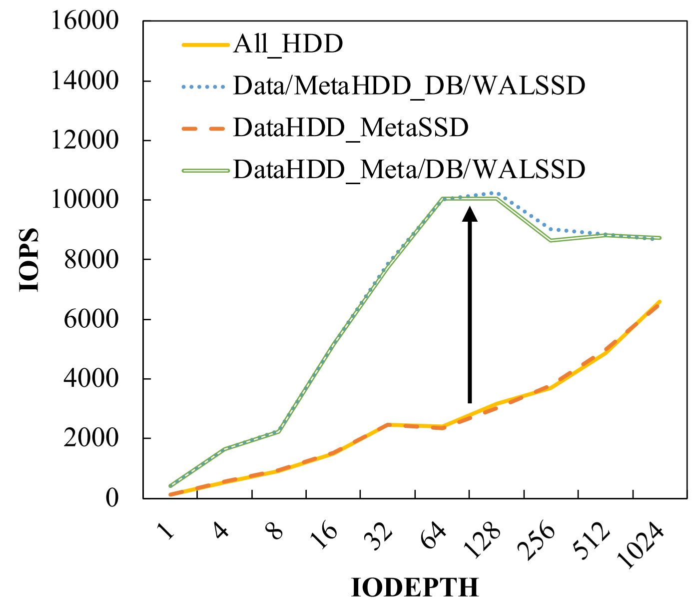
Yuki Kawada, Yoichiro Tanaka, Evaluation study of data access performance of distributed storage Ceph, using a brain neuronal structure visualization application, IEICE Tech. Rep., vol. 122, no. 63, MRIS2022-5, pp. 24-29, June 2022

Benchmark: DB/WAL stored in SSD

Write Bandwidth 2~6x

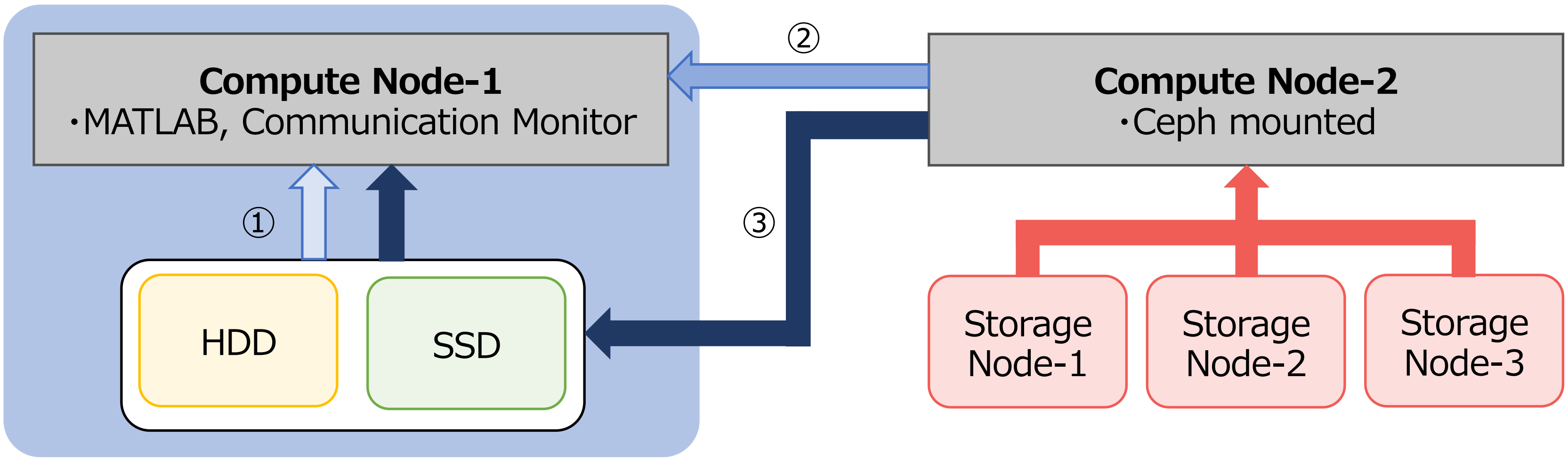


Random Write: IOPS up to 4x



Yuki Kawada, Yoichiro Tanaka, Evaluation study of data access performance of distributed storage Ceph, using a brain neuronal structure visualization application, IEICE Tech. Rep., vol. 122, no. 63, MRIS2022-5, pp. 24-29, June 2022

Retrieving Neuron Structure Data



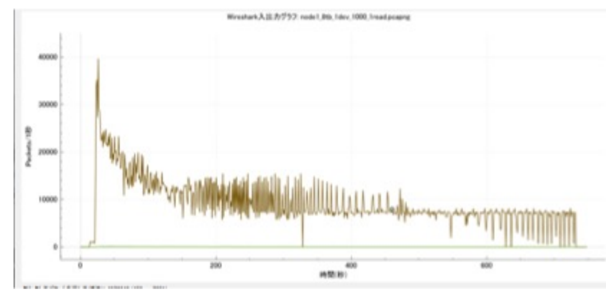
Data Retrieval Path

① Read directly from CN-1 device

② Read from CN-2-mounted Ceph

③ Copy data from CN-2 Ceph to CN-1, then read from CN-1 device

Dataset
1000frames: 10GB
5000frames: 50GB

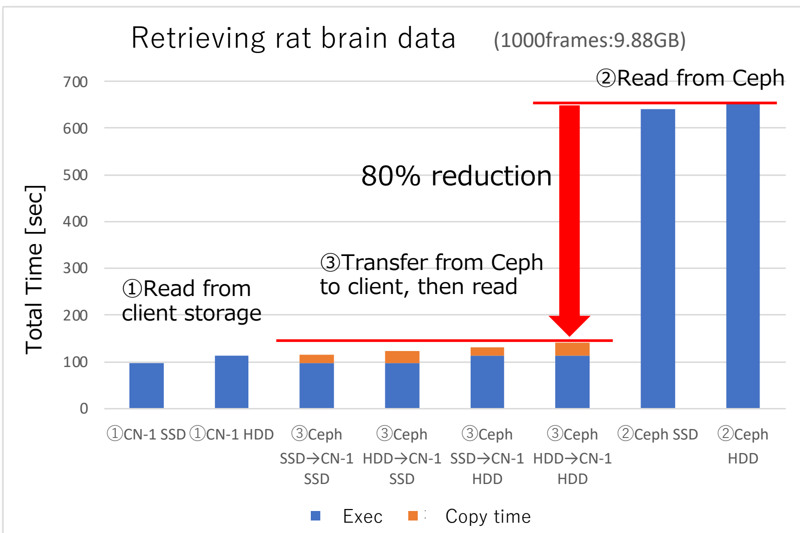


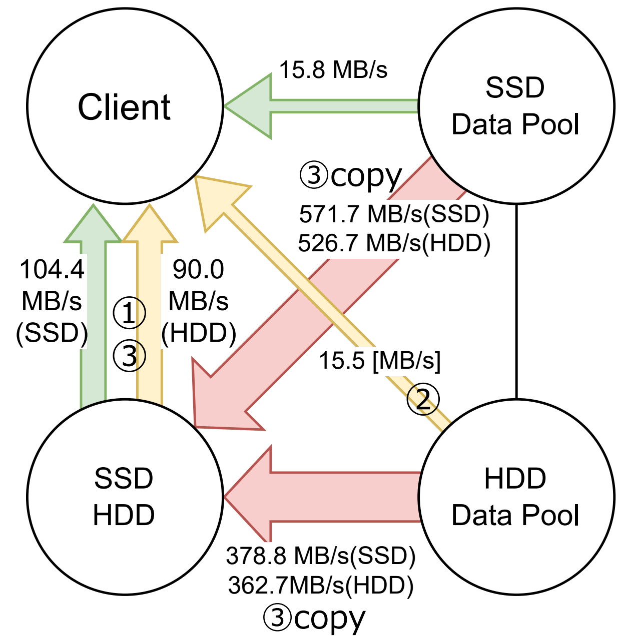
Communication

Yoichiro Tanaka
Research Institute of Electrical Communication, Tohoku University

Optimizing Data Path

Improve total data transaction efficiency by eliminating gating path in both compute and storage





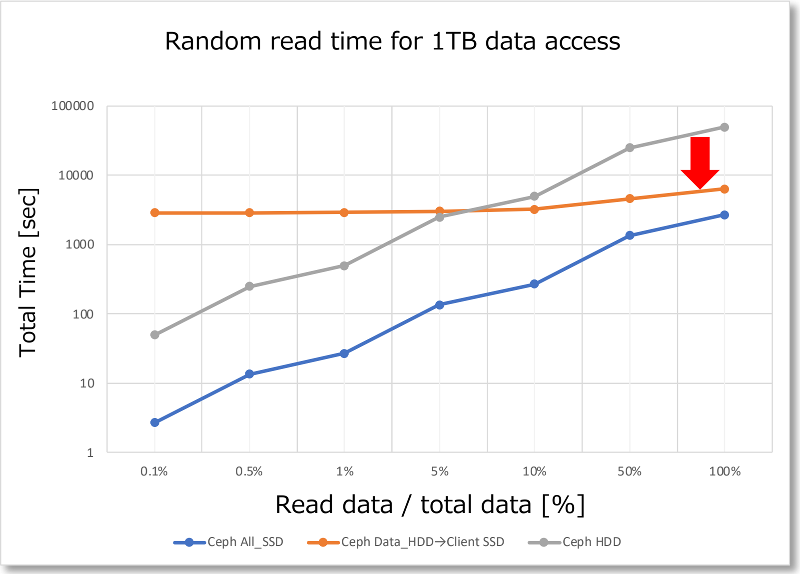
Exec	実行時間	96.9	112.4	96.9	96.9	112.4	112.4	641.2	651.6
Copy	コピー時間			17.7	26.7	19.2	27.9		
Total	合計時間	96.9	112.4	114.6	123.6	131.6	140.3	641.2	651.6

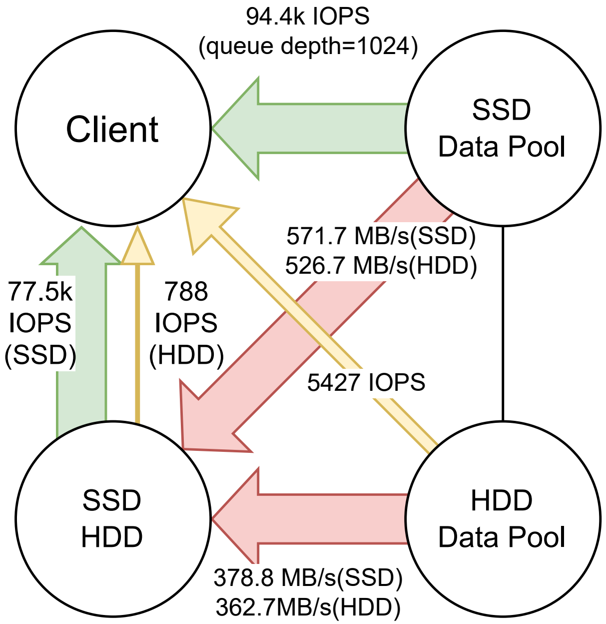
[sec]

Yoichiro Tanaka
Research Institute of Electrical Communication, Tohoku University

Optimizing Random Read

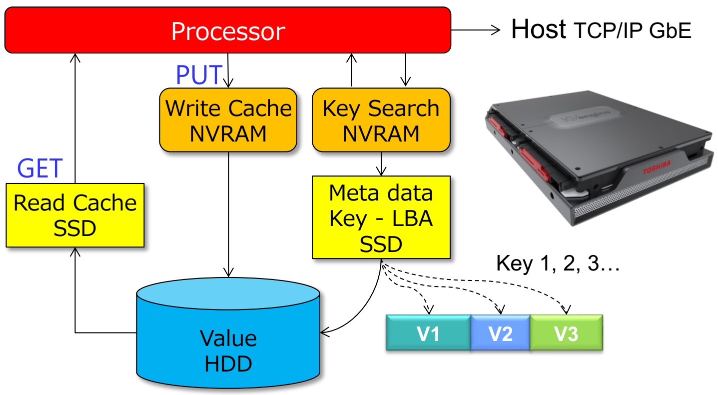
In case random read is more than 5% of total access data, total read time will be faster after transferring to client SSD

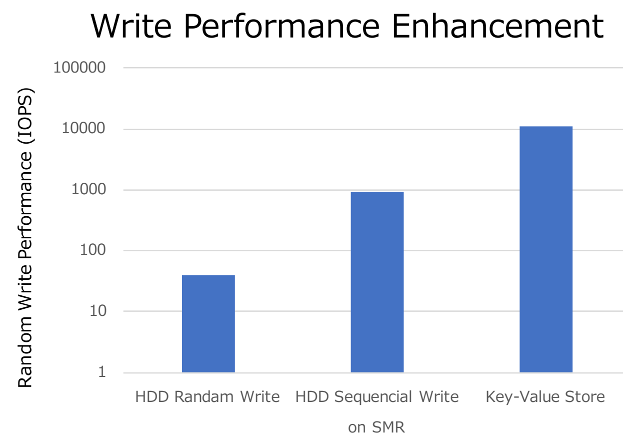


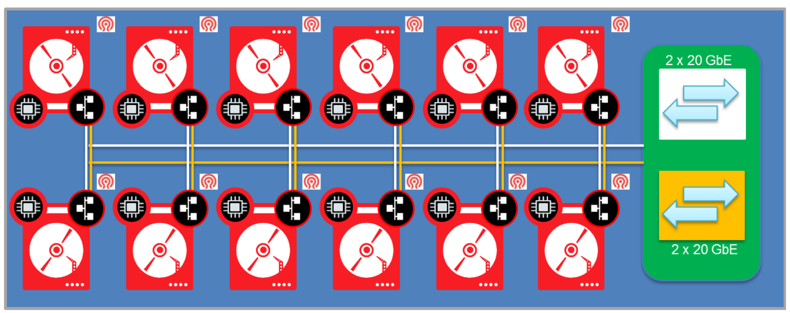


Yoichiro Tanaka
Research Institute of Electrical Communication, Tohoku University

Key-Value Store as an Atomic Module







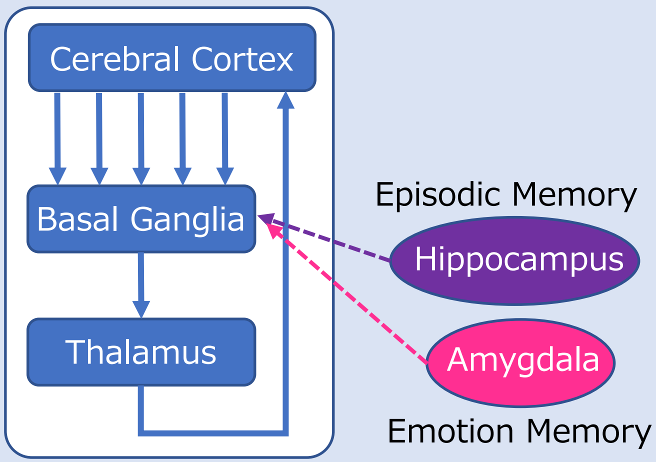
72 TB/12 Node Ceph Cluster in 1U Ethernet JBoD with KVS

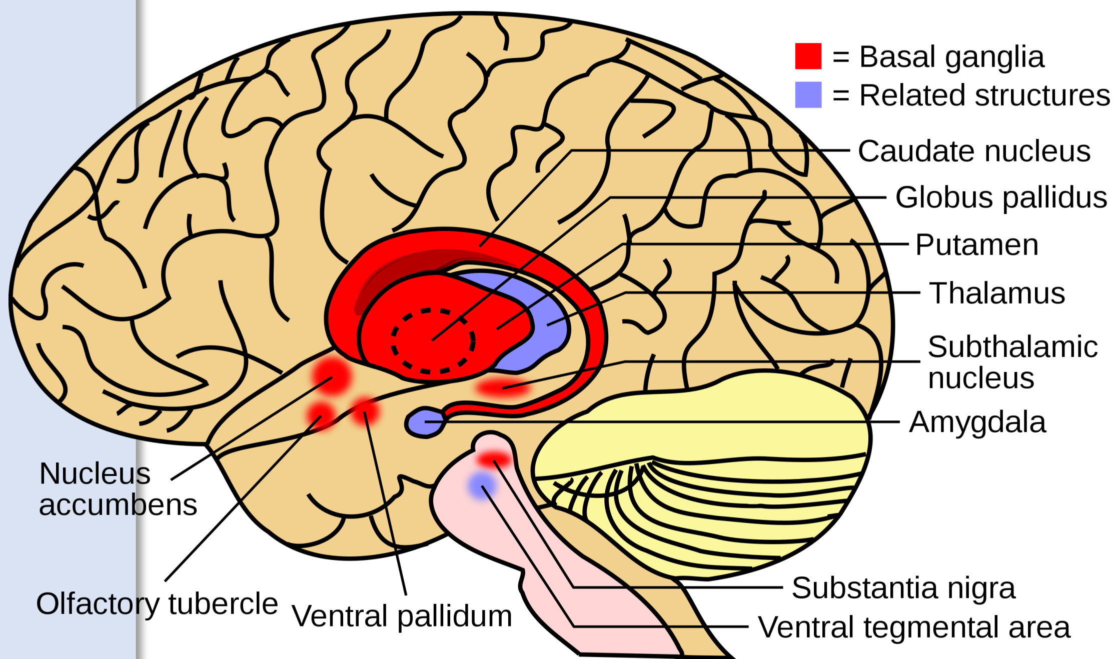
- Supermicro 1U/12 Drive Ethernet Chassis with 4 x10 GbE
- KVDrive with Ceph OSD – 12 Node Scale Out Cluster in a 1U
- Total of 72 TB of PMR Hard Drive Storage Capacity

S. Tanaka, M. Goto, and P. Kufeldt, KVDrive' Internet protocol drive for object storage systems, Toshiba Rev., Vol. 70, No. 8, 9–12, 2015
Toshiba Corp., Presented at Openstack Summit 2015, Vancouver, May 2015
Yoichiro Tanaka, Characterizing Advanced Recording Technology Assets with Hyperscale Applications, IEEE Trans. Magn., Vol. 52, No. 2, 3100404, 2016

Basal Ganglia as a Compute Module

Cortico-basal Ganglia Loop

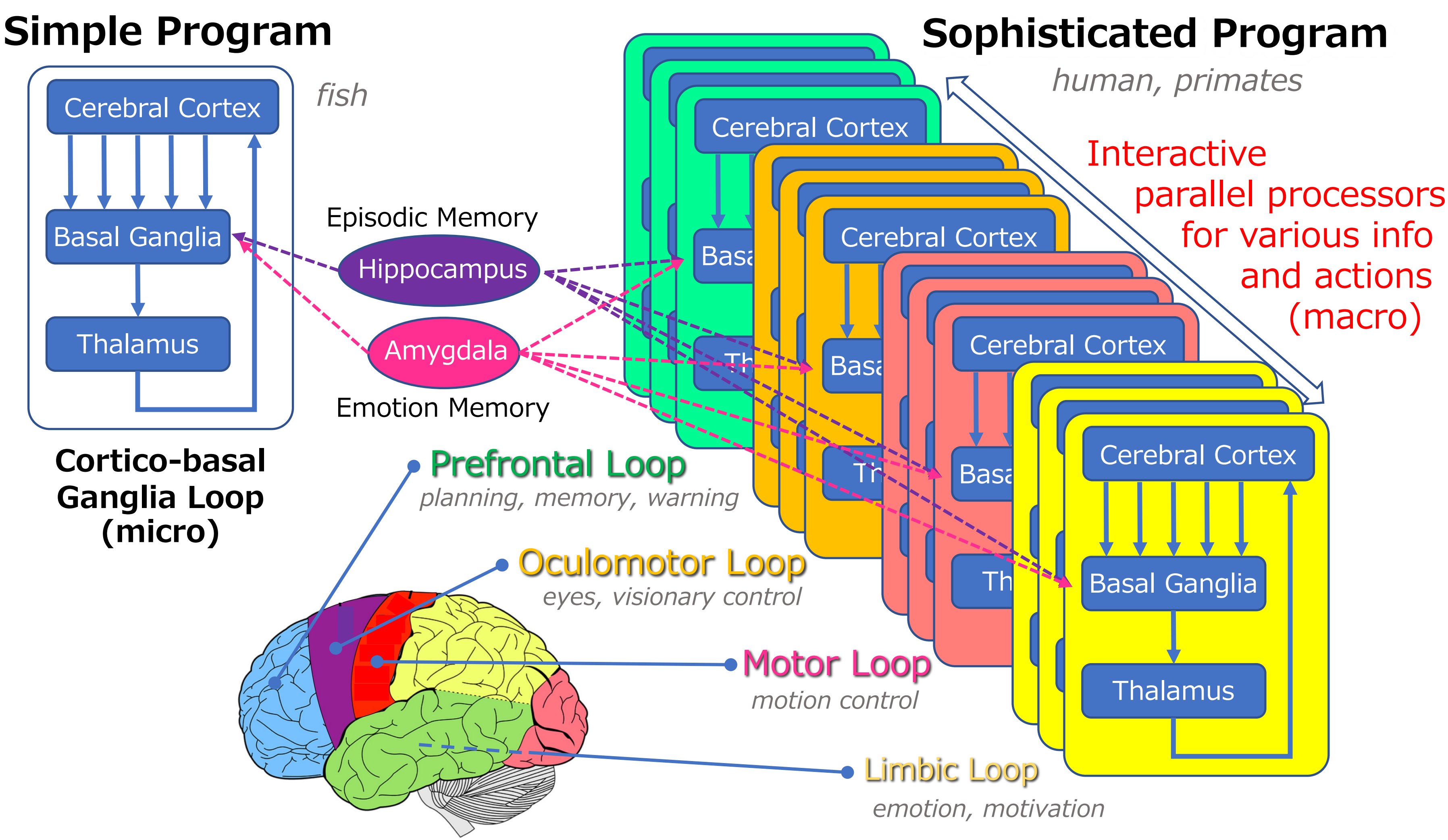




Simple Program Compute

- Long-term & Short-term Memory
- Information Path Selection
- Weighing Connections

Parallel Cortico-basal Ganglia Network



McHaffie JG1, Stanford TR, Stein BE, Coizet V, Redgrave P.(2005) Subcortical loops through the basal ganglia. Trends Neurosci. 28(8):401-407

Modularized Compute in Brain

In Cortico-basal Ganglia Loop;

- Long-term & Short-term Memory
- Information Path Selection
- Weighing Connections

There is NO digital bit stored.
There are weighing connections.

Modularized Computational Storage

- Key-Value data allocation in tiered memory stack for fast access to data
  - ✓ Long-term & Short-term Memory
  - ✓ Information Path Selection
  - ✓ Weighing Connections

Computing

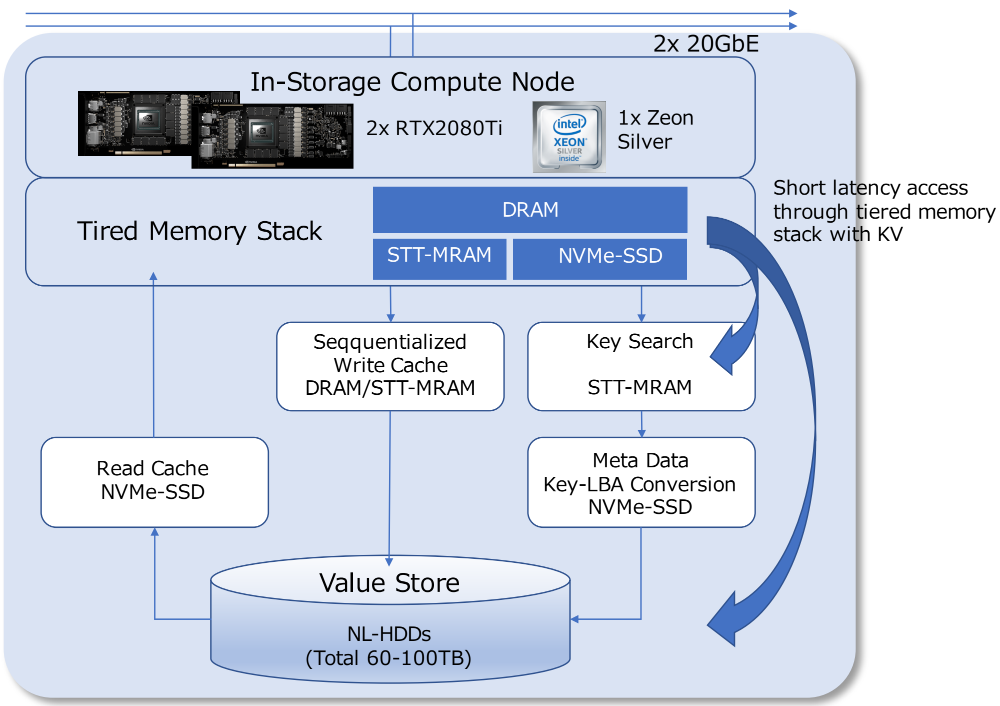
Distributed Analytics
I/F Control
Encryption & Blockchain

High Speed Key Search
Meta Data Search
Write Cache

Key-LBA Conversion
Read Cache

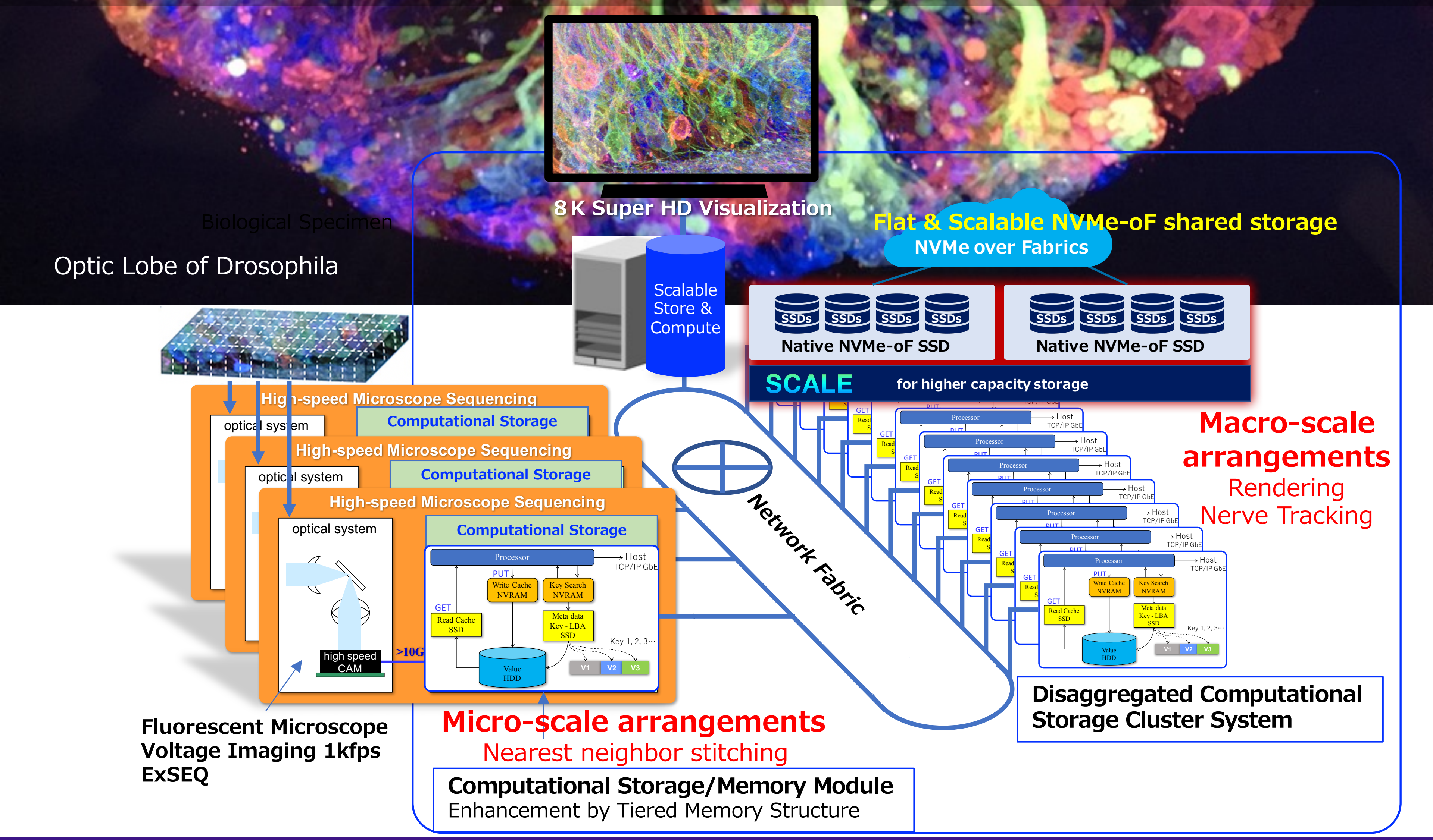
High Capacity Value Store

Non-Volatile Devices
(NV-Memory & Storage)



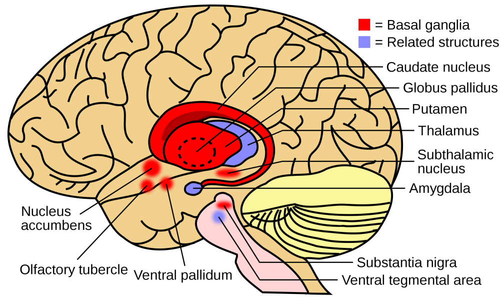
Yoichiro Tanaka, Characterizing Advanced Recording Technology Assets with Hyper-Scale Applications, IEEE Trans. Magn., Vol.52, No.2, pp.1-4, 2016
http://news.toshiba.com/press-release/business-and-retail-solutions/toshiba-demonstrates-high-performance-object-storage-tec, also presented at Openstack Summit 2015, Vancouver, May 2015

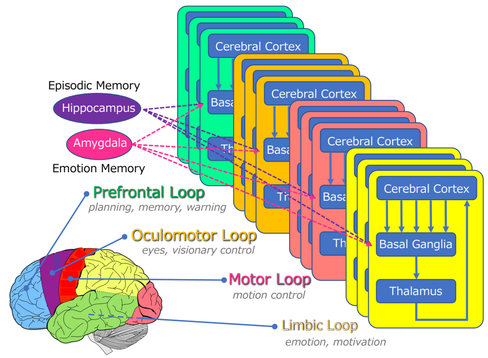
3D Visualization of Neural Structure inspired by brain functions

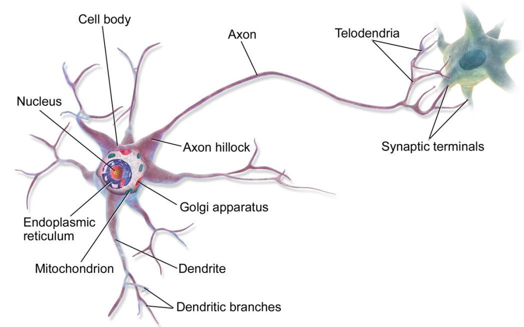


Yoichiro Tanaka
Research Institute of Electrical Communication, Tohoku University

Multi-scale Structures as a System







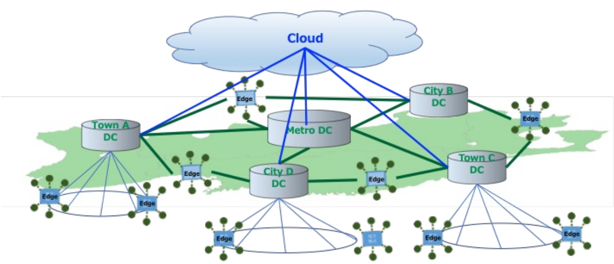
Node: *cortical region* — *cortical column* — *neuronal soma*

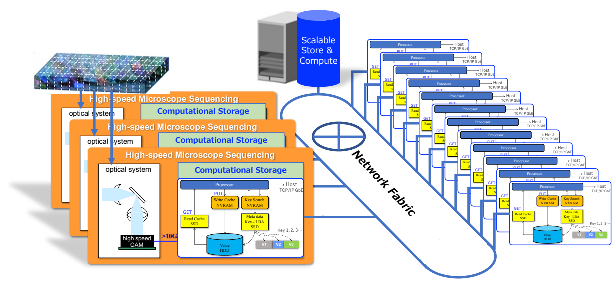
Edge: *white matter pathway* — *cortico-basal ganglia loop* — *neuron & synapse*

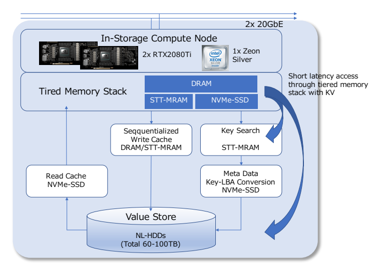
Macro-scale **Meso**-scale **Micro**-scale

Node: *cloud* — *disaggregated cluster* — *atomic module*

Edge: *local data center* — *computational storage system* — *integrated module*



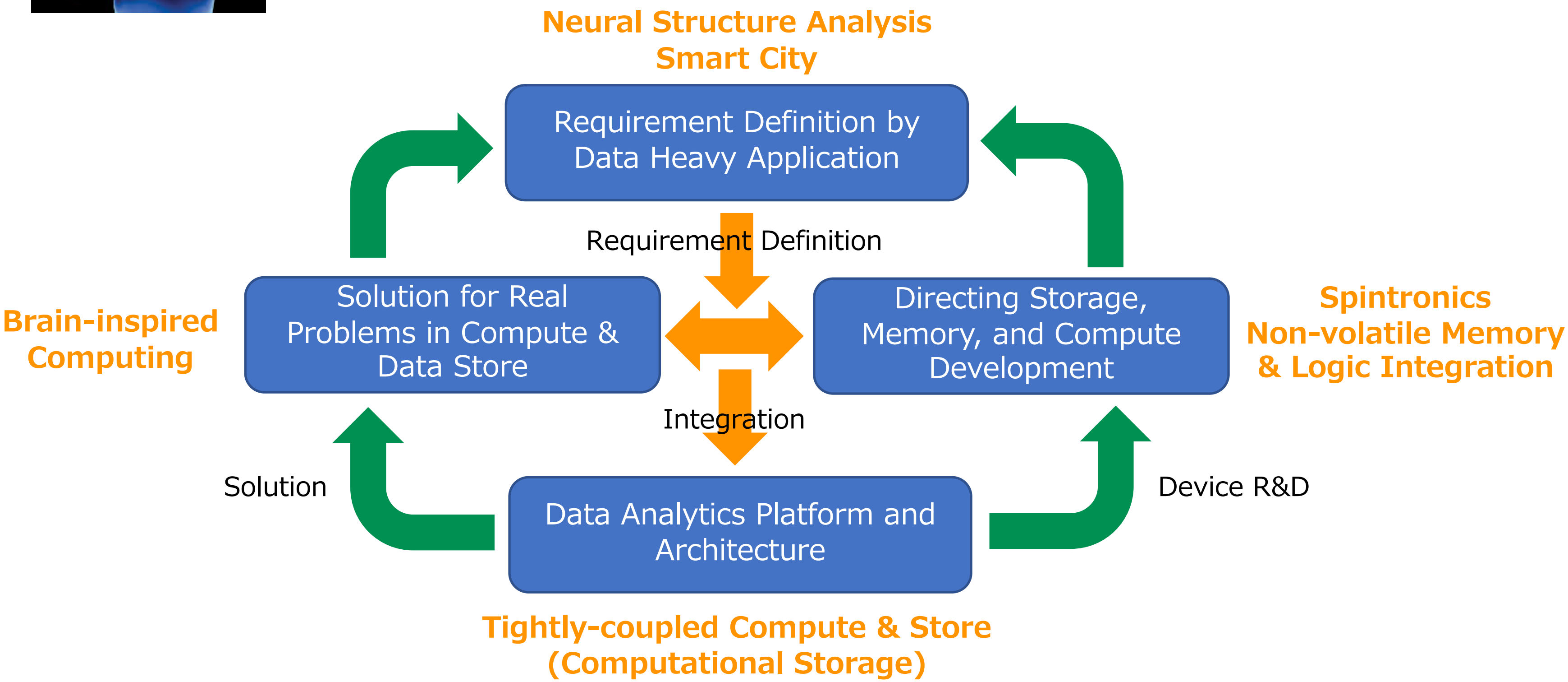




Yoichiro Tanaka
Research Institute of Electrical Communication, Tohoku University

Data Store and Brain

Re-define "data storage" as a computation-unified active system by learning from brain functions



Summary

1. Perpendicular magnetic recording has created the foundation of digital data society. The storage technology provides the most significant contribution to our society in ICT and Big Data.
2. In life science analytics, handling large scale real-time data analysis and secure management of unstructured datasets are important. Unification of compute and storage closely to data source is required.
3. Parallel processing in the “Cortico-basal Ganglia Loop” module structure is a good references for modularized computational storage scheme. This is scalable and applicable to large scale dataset analytics.

Acknowledgements

Great thanks to research collaboration with

- Dr. Y. Bando, Synthetic Neuroscience Team, MIT Media Lab
- Prof. A. Hirano, Assoc. Prof. H. Yamamoto, Assoc. Prof. Simon Greaves, Yuki Kawada, Tohoku University

Part of the research is supported by;

- The Cooperative Research Project Program of the Research Institute of Electrical Communication, Tohoku University
- JSPS Grant-in-Aid for Scientific Research (B) 20H02194
- Tohoku University Advanced Institute of Yotta Informatics
- Kioxia Corporation
- HGST Japan (Western Digital Japan Corporate)