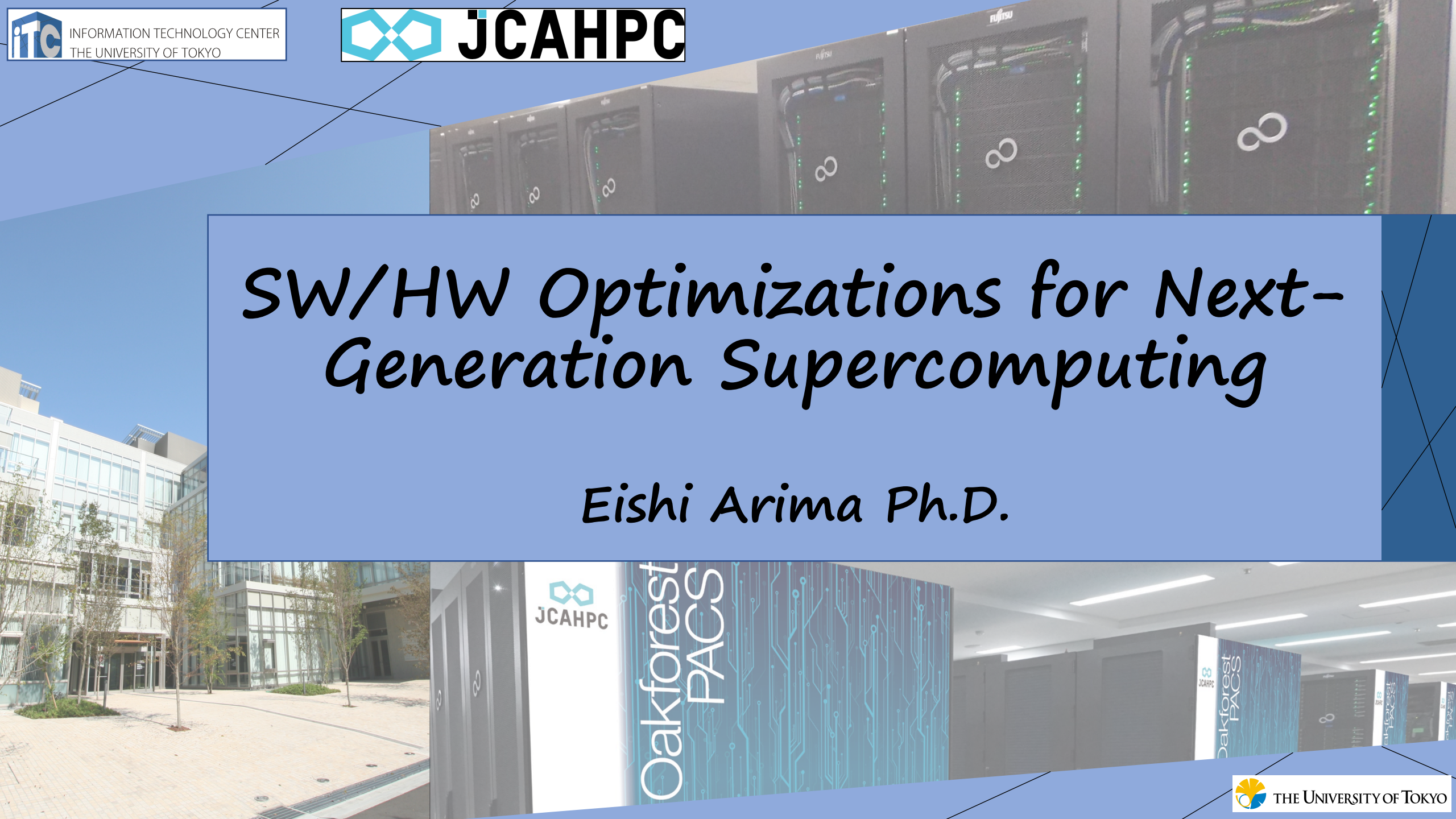
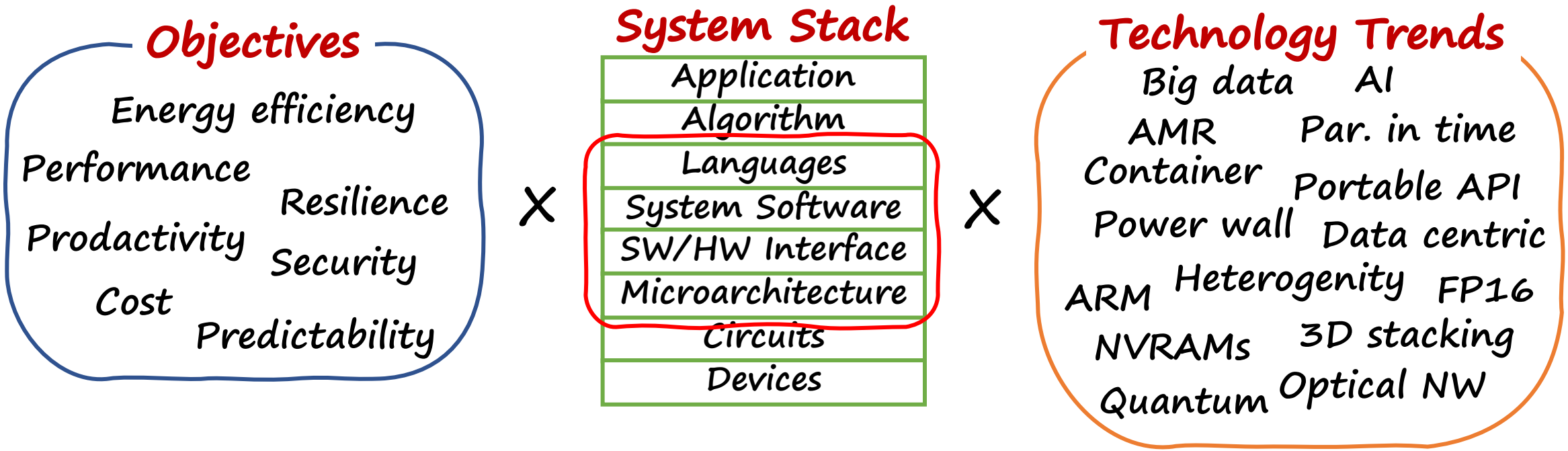

INFORMATION TECHNOLOGY CENTER
THE UNIVERSITY OF TOKYO
JCAHPC
SW/HW Optimizations for Next-Generation Supercomputing
Eishi Arima Ph.D.
JCAHPC
Oakforest PACS
THE UNIVERSITY OF TOKYO

Research landscape & our focus

Goal: to improve performance and energy efficiency of current/future supercomputing systems

Approaches: from microarchitecture to programming models; esp. interactions among them

Key words: power/data managements; heterogenity; new HW devices

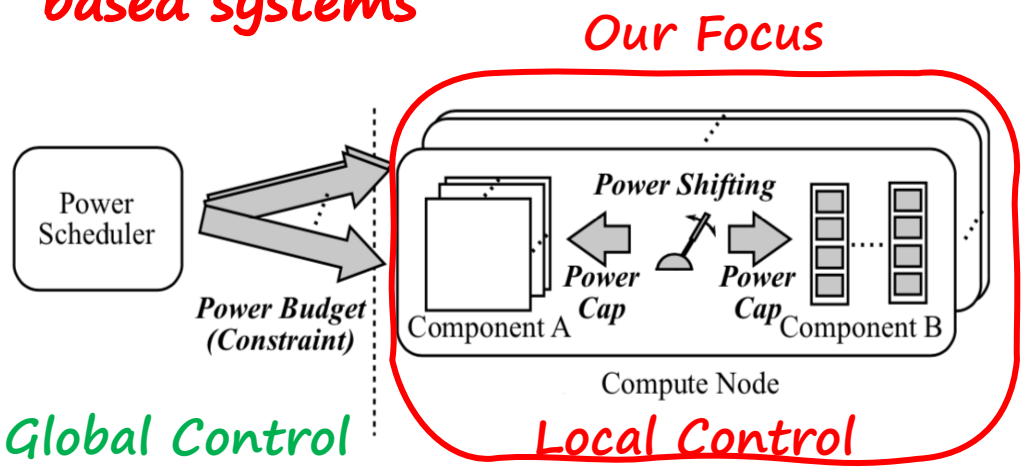


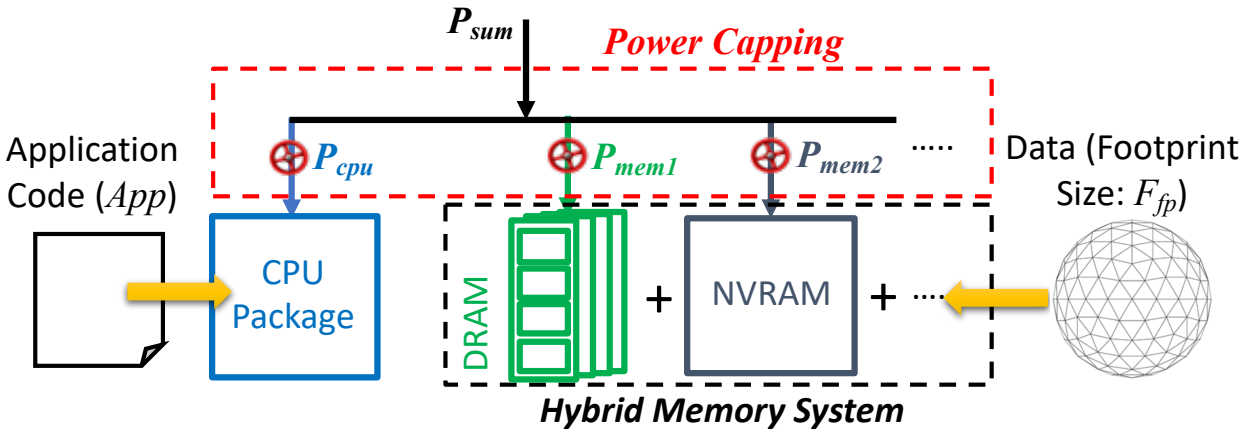
Footprint-Aware Power Capping [E. Arima+ ISC'20]

Background: Hierarchical power management will be an indispensable feature for future power-constrained supercomputers

- **Global control:** Power scheduler distributes power budgets accordingly to all compute nodes
- **Local control:** Each compute node attempts to exploit the allocated power budget

Our insight: Optimal power allocations to components can change considerably when the problem size (= **data footprint size**) is scaled esp. on **hybrid memory based systems**





Footprint-Aware Power Capping [E. Arima+ ISC'20]

1st Contribution: problem formulation & performance modeling

2nd Contribution: a profile-based software framework (efficient calibration method + power allocation algorithm)

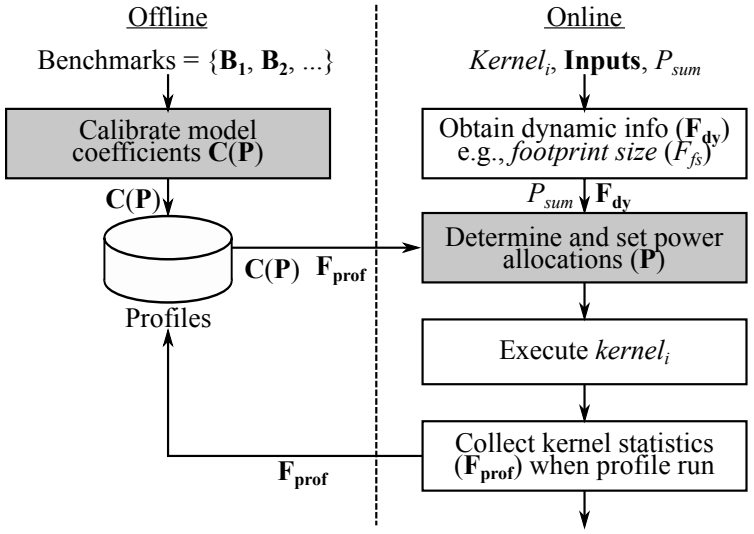
Result: near optimal allocations acquired!

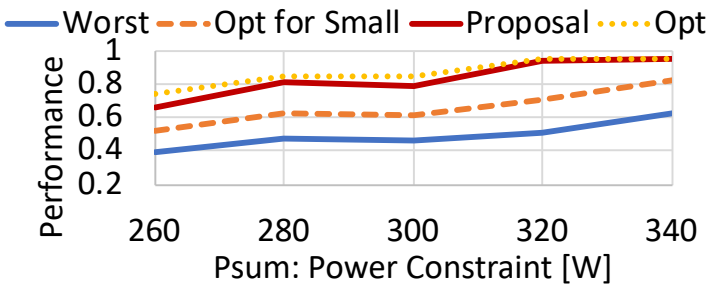
Given *Kernel*, **Inputs**, P_{sum} ($\Longrightarrow$ **F**, P_{sum})
Max *Obj*(**P**, **F**)
s.t. $\Sigma P_x \leq P_{sum}$
$P_x \in S_{Px}$ (x = cpu, mem1, …)

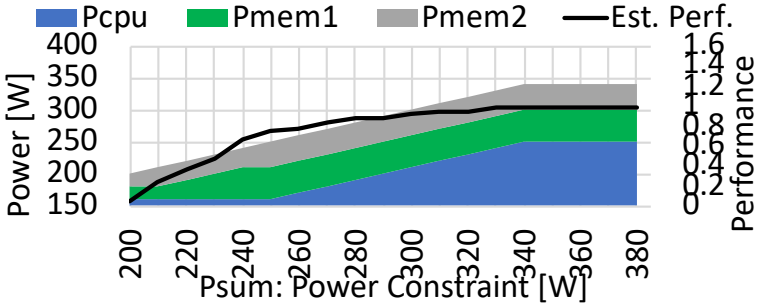
Kernel: target kernel region
Inputs: inputs for the app = (arg1, arg2, …)
F: app feature parameters (e.g., F/B rate)

Obj(**P**, **F**): objective function

P_{sum}: given total power budget [W]
P: power allocations = (P_{cpu}, P_{mem1}, …)
S_{Px}: Set of power caps for component *x* (e.g., ={20, 30, 40})







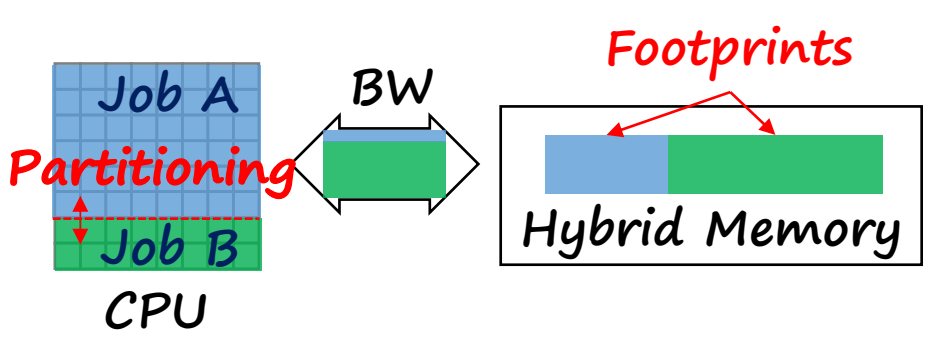
Next Step: Scheduler-Based Appproaches (Ongoing Work)

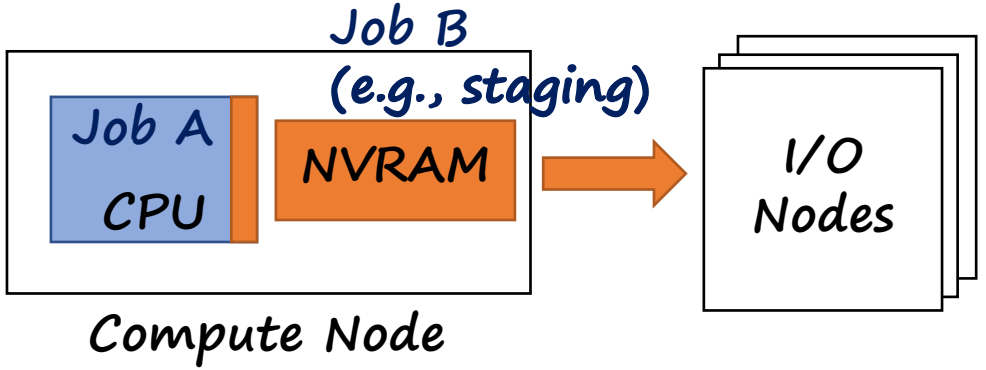
Objective: developing job/power scheduling methods using the **insights** gained through the **power capping work**

- Footprint awareness; perfomrance prediction; system design
- Supported by JSPS grant-in-aid

Our focus: **co-scheduling, i.e., co-running multiple jobs within a node to improve total job throughput or energy efficiency**

- Footprint-aware co-scheduling [ICPP'19 poster] – ongoing
- I/O-aware co-scheduling [Not appeared yet] – ongoing

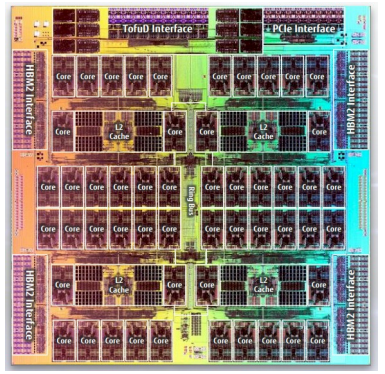




Power Evaluation using Supercomputer Fugaku

- Evaluation of Power Controls on Supercomputer Fugaku [Y. Kodama+ EEHPC@CLUSTER'20]
  - Large-scale power evaluation using **over 20K nodes** of the Fugaku system
  - Particularly, focusing on some power control features implemented on A64FX processor
- Evaluation for future HPC microprocessors
  - by extending existing perfororormance/power simulators used for the development of the Fugaku system (ongoing work)

Supercomputer Fugaku†



A64FX*

†https://www.riken.jp/en/news_pubs/news/2020/20200623_1/

*https://www.nextplatform.com/2019/11/13/a64fx-arm-chip-gets-a-big-push-from-cray/

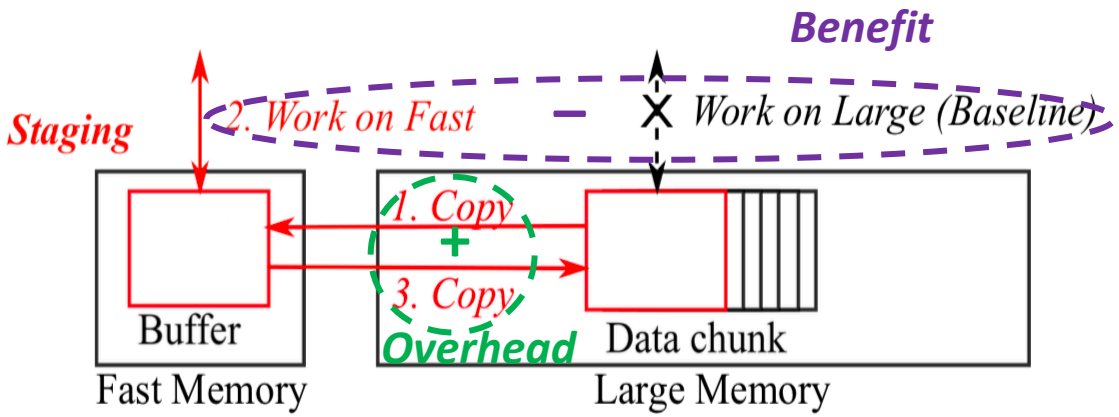
Pattern-Aware Staging [E. Arima+ISC'20]

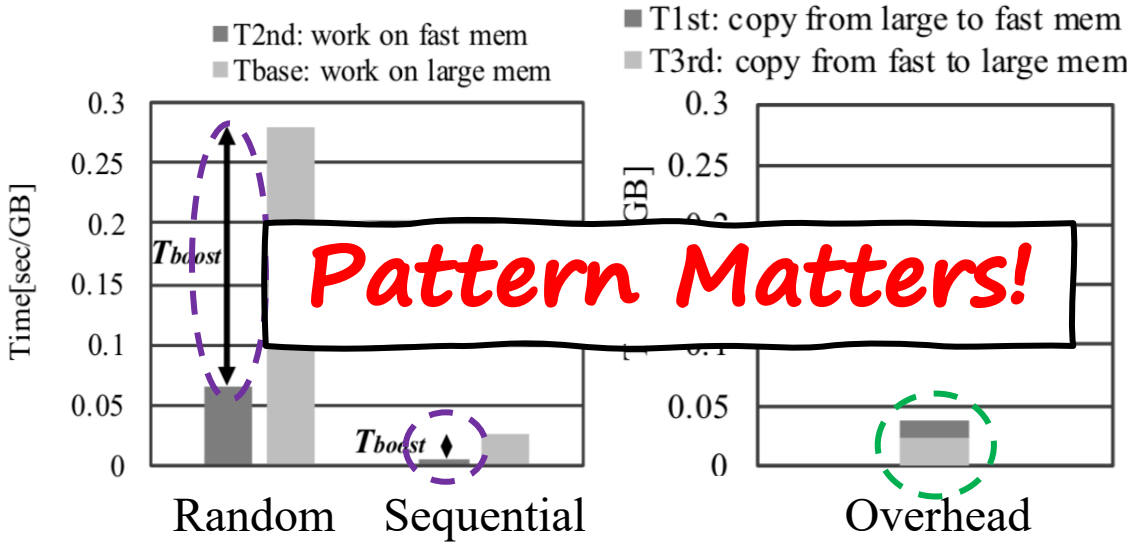
Open question: How can we exploit both the high BW and the large capacity on a hybrid memory system?

- E.g., HBM + DDR4 or DRAM + NVRAM

Observation: The access pattern matters when we decide whether or not move a chunk of data

- Performance benefit v.s. data movement overhead



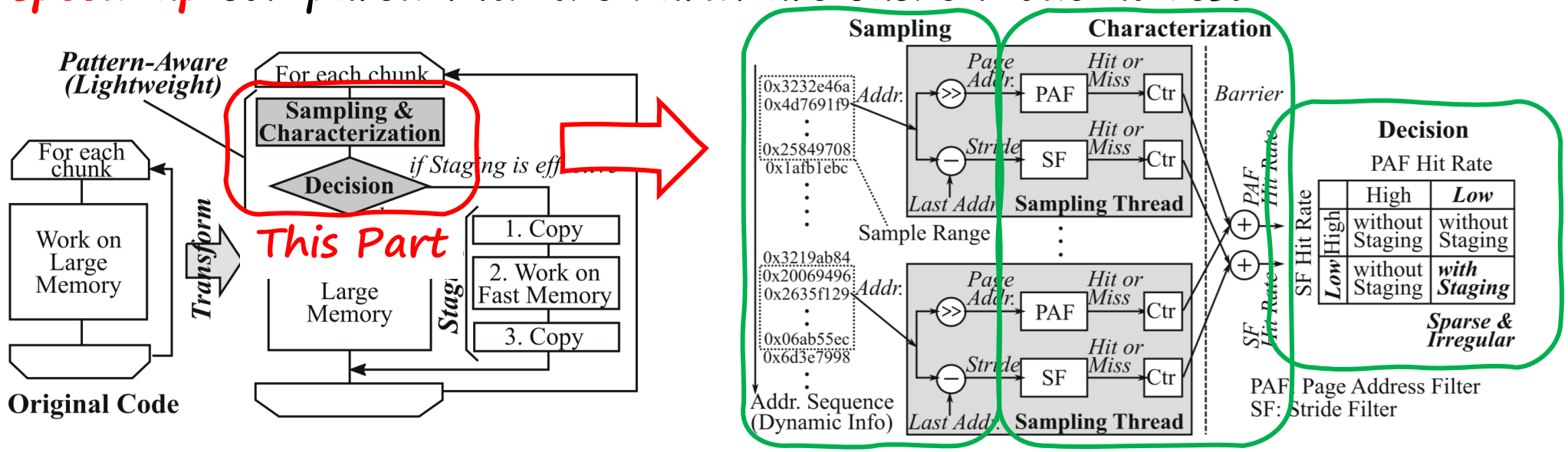


Pattern-Aware Staging [E. Arima+ISC'20]

Pattern-aware staging: a lightwieght software mechanism that can be ultimately automated by source-to-source compilers

1. **Sampling:** helper threading inspired dynamic address sequence sampling
2. **Characterization:** Bloom filter based runtime access pattern detection
3. **Decision:** criterium to determine whether or not move a chunk of data

Evaluation results: **300% speed-up** compared with the large memory only and **41% speed-up** compared with the hardware cache mode at best

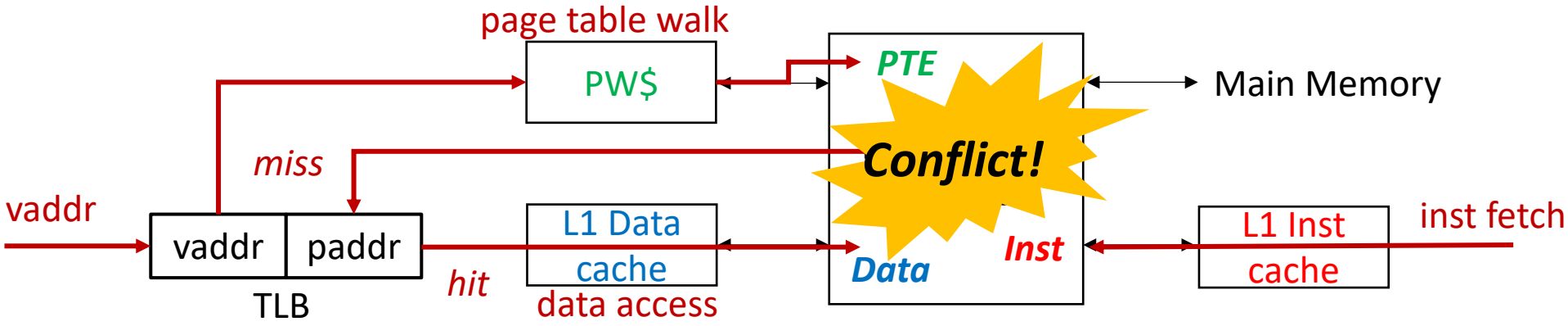


Advanced HW Cache Management [E. Arima DSD'20]

Background: significant cache conflicts among different classes of cachelines on lower level caches

- Different classes: data, instructions (codes), and PTEs
- Even worse because: (1) PTE accesses are becoming critical on modern systems esp. for virtual machines; (2) code footprints are increasing esp. for modern server workloads

We should set appropriate allocation priorities to these classes!



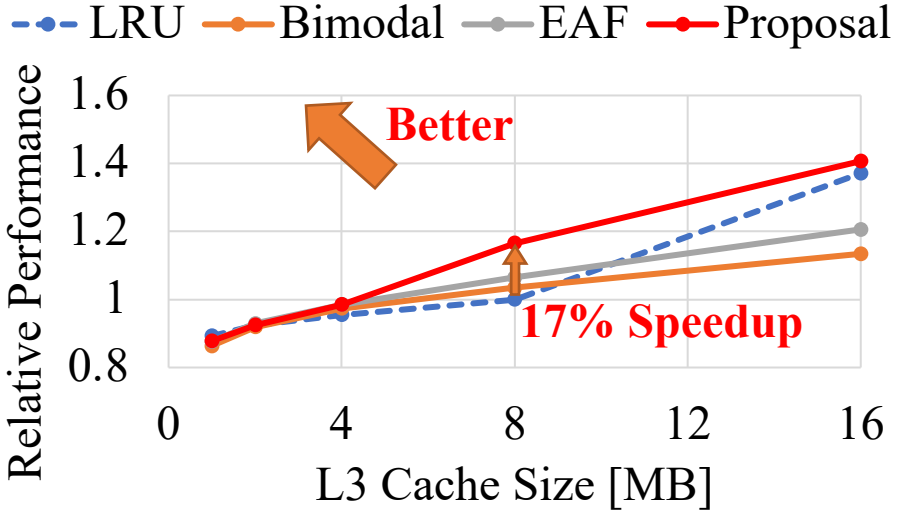
Advanced HW Cache Management [E. Arima DSD'20]

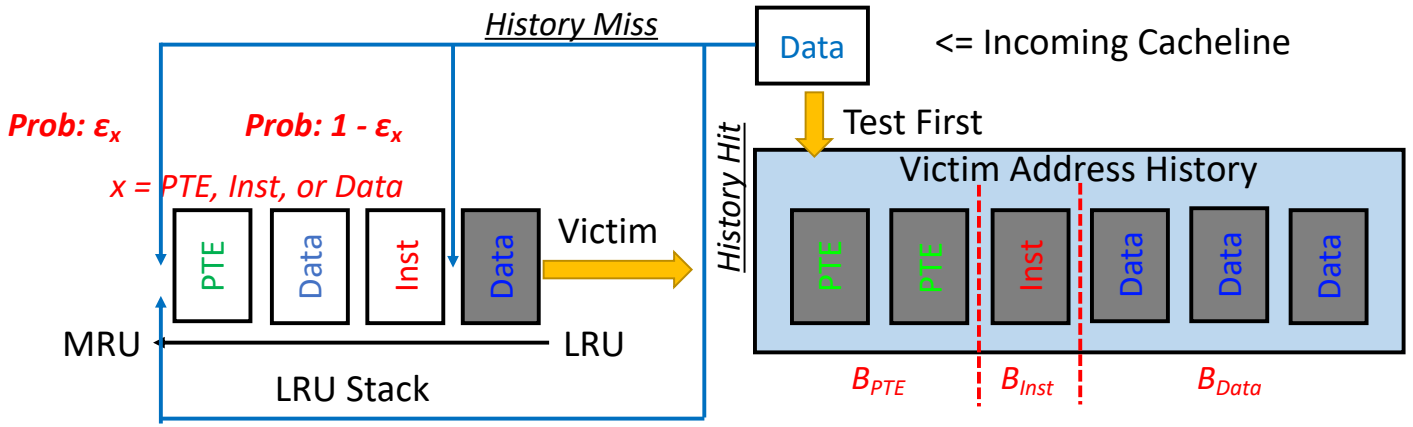
Approach: (1) augment a functionality to set a unique priority to each class;

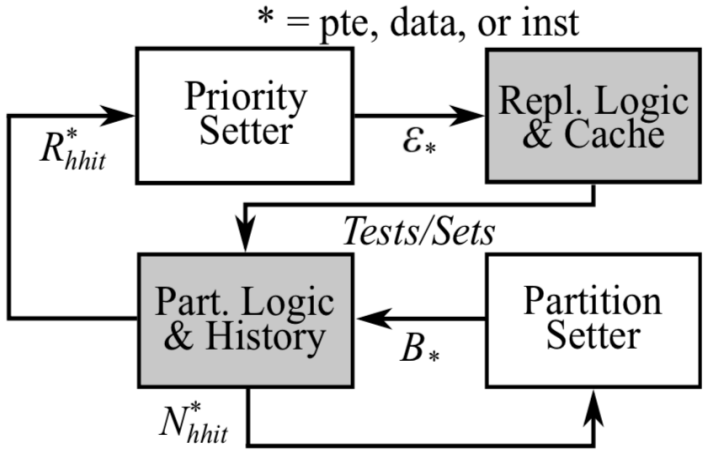
(2) partition a module used for estimating reuseness behavior (history of victims);

(3) provide a control system to optimize the **priorities/partitions at the same time**

Evaluation result: considerable speed-up over existing approaches







Conclusion

Summary: We are aiming to provide various **insights** that can potentially lead to significant performance, energy efficiency, or any other sorts of improvement for current/future supercomputing systems through both SW & HW based approaches.

Related Publications:

- E. Arima, T. Hanawa, C. Trinitis, M. Schulz "Footprint-Aware Power Capping for Hybrid Memory Based Systems" In *Proc. of ISC High Performance*, pp.347-369, (2020) Youtube
- Y. Kodama, T. Odajima, E. Arima, and M. Sato "Evaluation of Power Controls on Supercomputer Fugaku" In *Proc. of CLUSTER (EEHPC volume)*, pp.xx-xx, (2020) Youtube
- E. Arima, M. Schulz "Pattern-Aware Staging for Hybrid Memory Systems" In *Proc. of ISC High Performance*, pp.474-495, (2020) Youtube
- E. Arima "Classification-Based Unified Cache Replacement via Partitioned Victim Address History" In *Proc. of DSD*, pp.101-108, (2020)

🌐 **Website:** https://www.cspp.cc.u-tokyo.ac.jp/arima/index-e.html

✉ **E-mail:** arima@cc.u-tokyo.ac.jp

That's it! Thank you for watching!

free sound effects from https://www.fesliyanstudios.com

Staff

Director: Eishi Arima

Speaker: Eishi Arima

Investigator: Eishi Arima

Editor: Eishi Arima

Grants

JSPS Grant-in-Aid for Early-Career Scientists (JP20K19766) – PI: EA

JSPS Grant-in-Aid for Early-Career Scientists (JP18K18021) – PI: EA

JSPS Grant-in-Aid for Research Activity Start-up (JP16H06677) – PI: EA

Research on Processor Architecture, Power Management, System Software and Numerical Libraries for the Post K Computer System of RIKEN

Special Thanks to

Folks at ITC, U. Tokyo; CAPS, TUM; R-CCS, RIKEN; and CASC, LLNL

The End