



**Research On Power and Energy  
Efficiency of Supercomputers  
~ collaborated with LLNL and Titech ~**

---

Hiroshi Nakamura

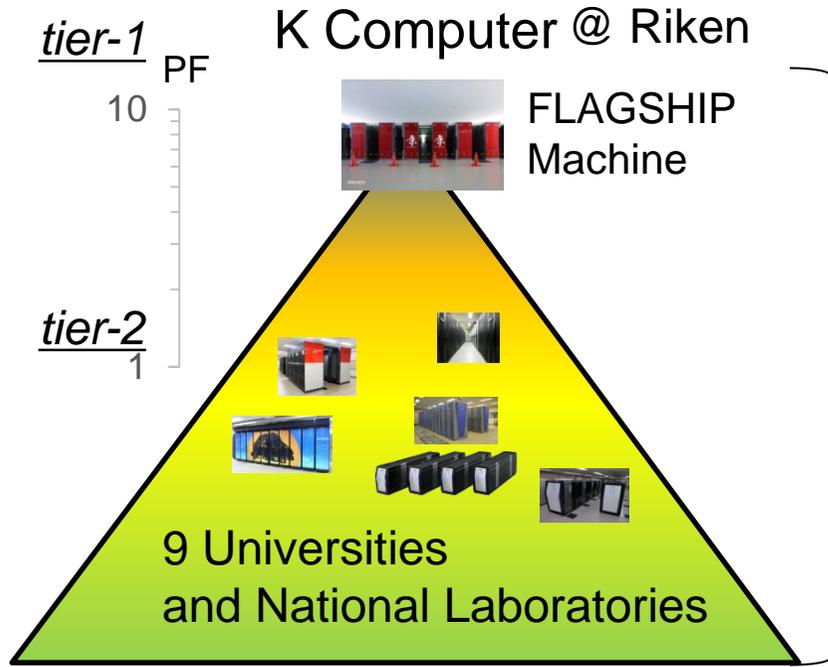
*Director of Information Technology Center,  
The University of Tokyo*

# Agenda



- Overview of Current Status in Japan
  - HPCI: High Performance Computing Infrastructure
- Research on Power and Energy Efficiency of Supercomputers at UTokyo
  - Power Steering for Over-provisioned System collaboration with Titech and LLNL

# Overview: Supercomputers in Japan



These supercomputers form **HPCI**  
(High Performance Computing Infrastructure)

Features:

- Single sign-on
- Shared storage (Distributed file system)

## Joint Usage/Research Center

for Interdisciplinary Large-scale for information Infrastructures

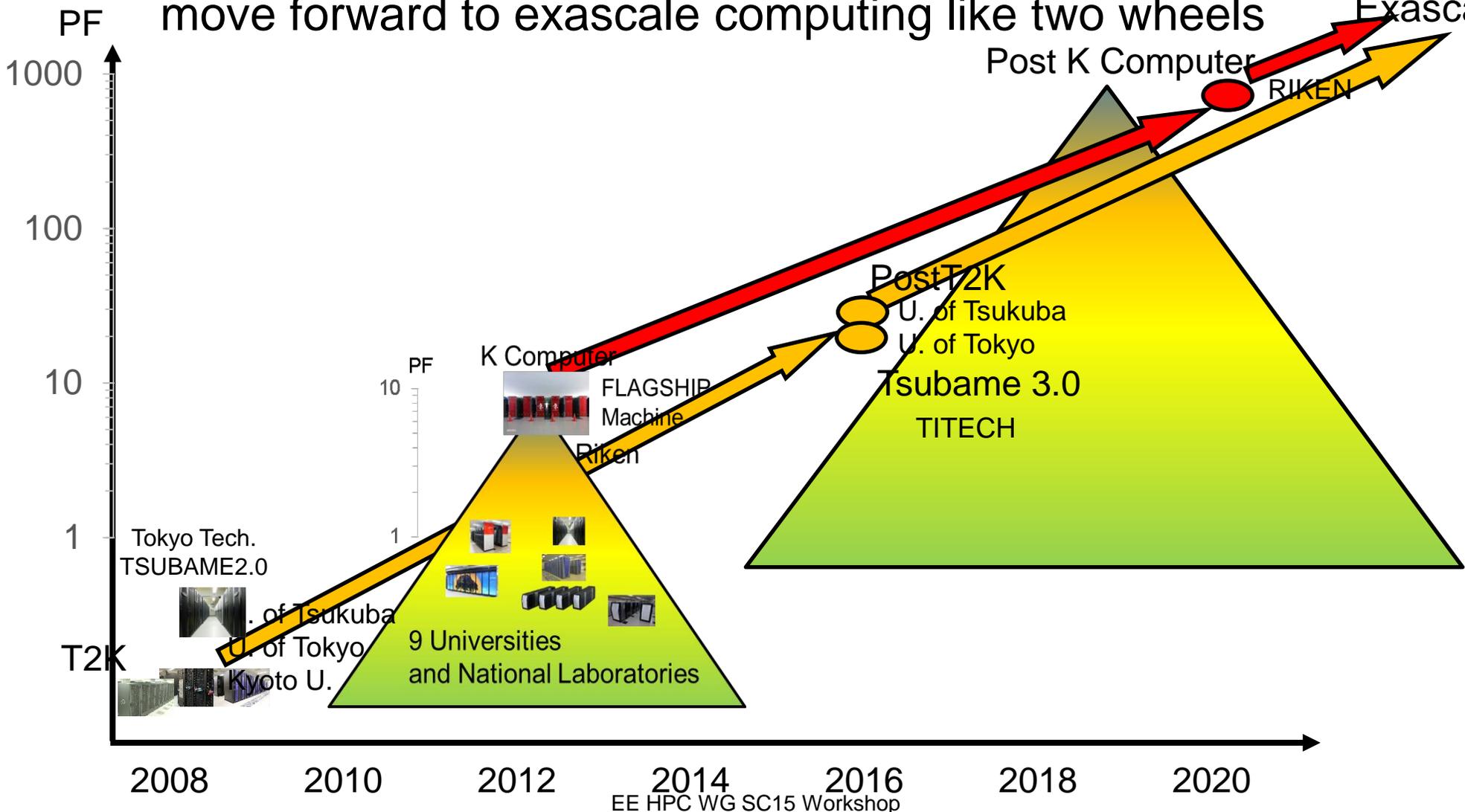
- network-type: eight of tier-2 Universities
- headquarter: University of Tokyo
- promote academic research through synergy of eight centers



# Towards Exascale Computing

tier-1 and tier2 supercomputers form HPCI and move forward to exascale computing like two wheels

Future Exascale



# Agenda



- Overview of Current Status in Japan
  - HPCI: High Performance Computing Infrastructure

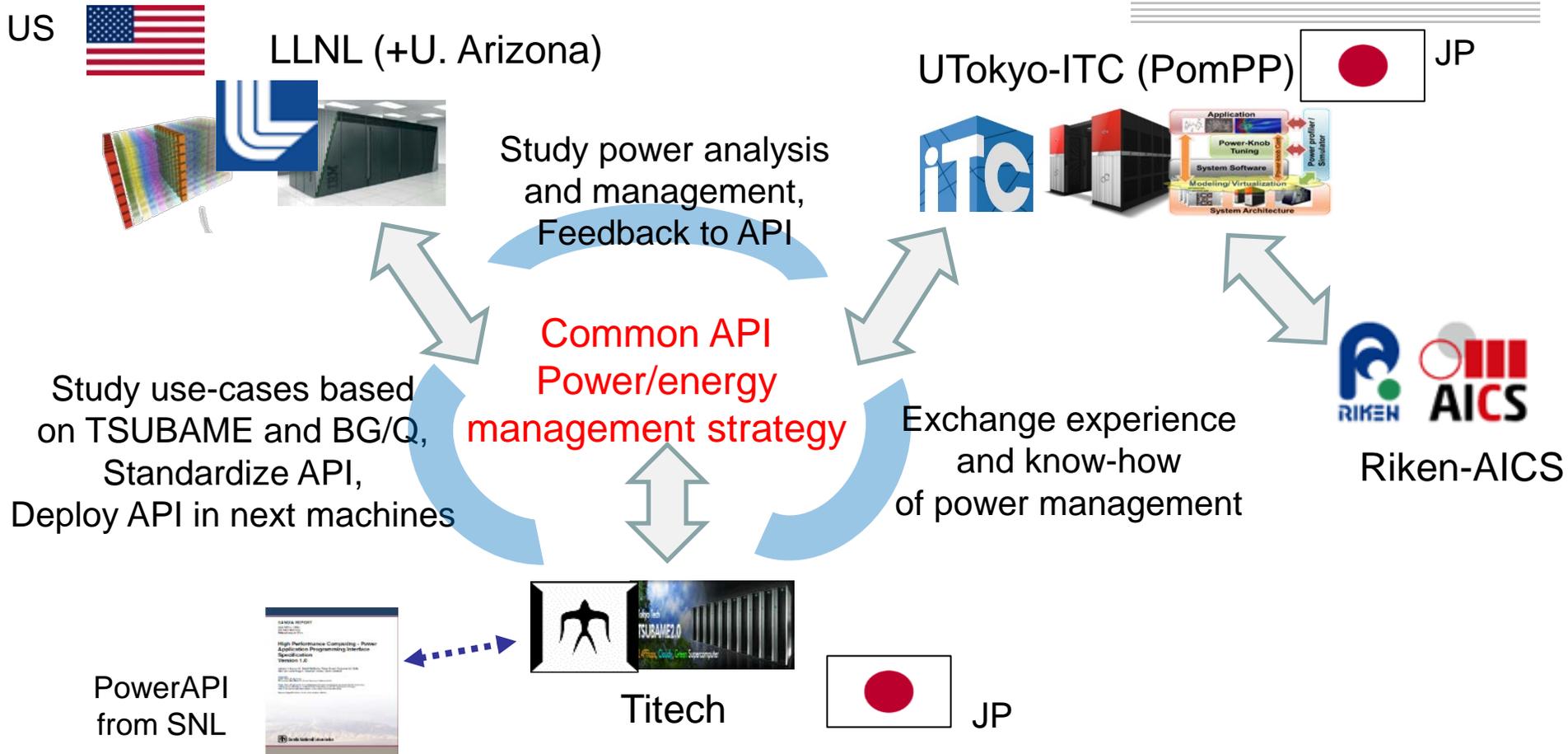
- 
- Research on Power and Energy Efficiency of Supercomputers at UTokyo
    - Power Steering for Over-provisioned System  
collaboration with Titech and LLNL

# US DoE / JP MEXT Open Scientific Research Collaboration



- Purpose
  - Work together where it is mutually beneficial to expand the HPC ecosystem and improve system capability
- Joint Activities
  - Pre-standardization interface coordination
  - Collection and publication of open data
  - Collaborative development of open source software
  - Evaluation and analysis of benchmarks and architectures
  - Standardization of mature technologies
- Technical Areas of Cooperation
  - Kernel System Programming Interface, Low-level Communication Layer Task and Thread Management to Support Massive Concurrency, **Power Management and Optimization**, Data Staging and Input/Output (I/O) Bottlenecks, ...

# Collective Plan (UTokyo-LLNL-Titech)



**Final goal: deploy common API and power management strategy in future extreme-scale systems**

# Current Activities

## System Oriented API

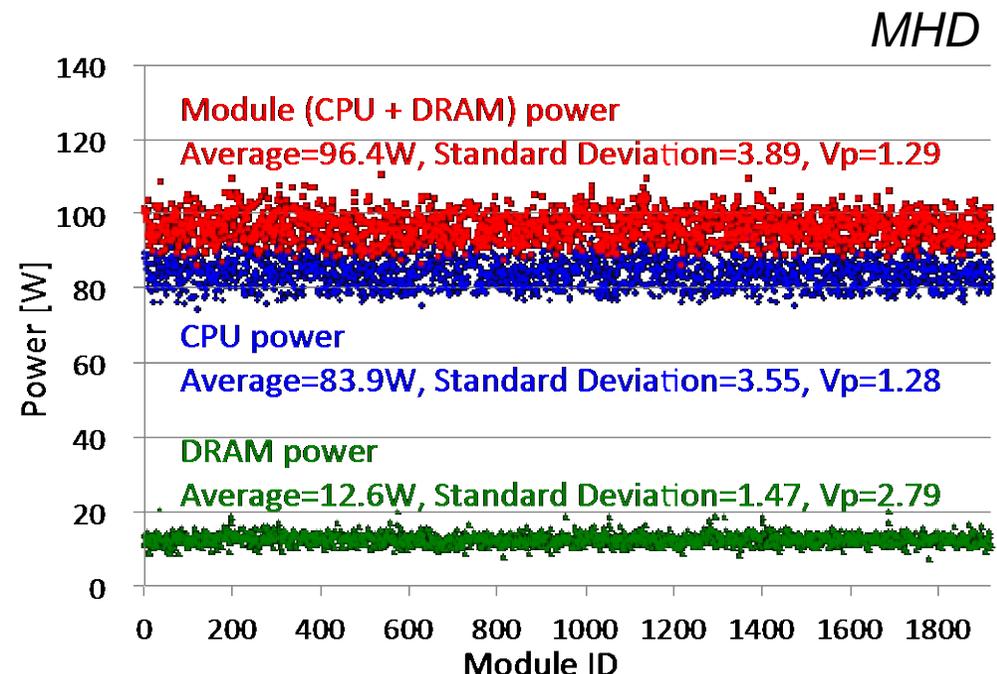
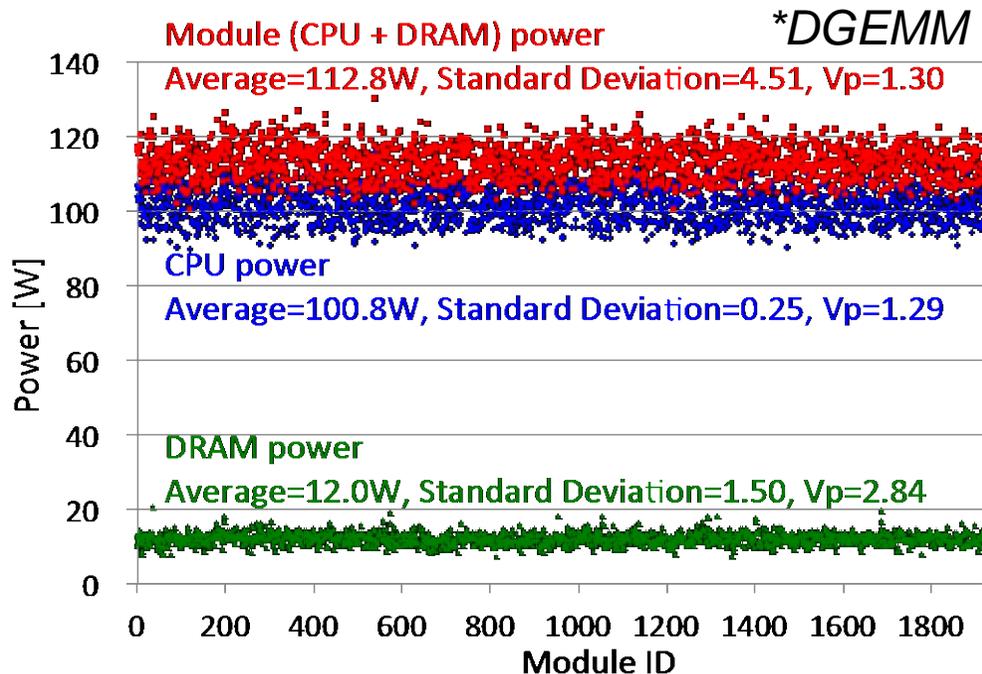
- Discussing system specifications related to power monitoring and control for future consideration of power APIs
  - Preparing a pilot system TSUBAME-KFC at GSIC Center, Titech
  - Collect experimental data for future discussion

## Power Steering Technique for Over-provisioned System

- Investigating a CPU module-wise power-budgeting strategy
  - SC 15 paper In cooperation with Kyushu U. and U. of Arizona  
[Y. Inadomi, et al., “Variation-Aware Power Budgeting in Power-Constrained High-Performance Computing”, SC’15, 2015. \(3:30 PM on Thu. at 19AB\)](#)
- Collaborated development of tools for power and performance prediction for interconnection networks
  - [S. Miwa, and H. Nakamura, “Profile-based Power Shifting in Interconnection Networks with On/Off Links”, SC’15, 2015 \(11:00 AM on Wed. at 18CD\)](#)

# CPU Module-Wise Power-Budgeting Strategy

- Motivation: Average power of CPU and Memory for each node
- Experimental environment: HITACHI HA8000-tc/HT210
  - Intel Xeon E5-2697 v2 (12 cores, 2.7GHz) x 2, 960-nodes



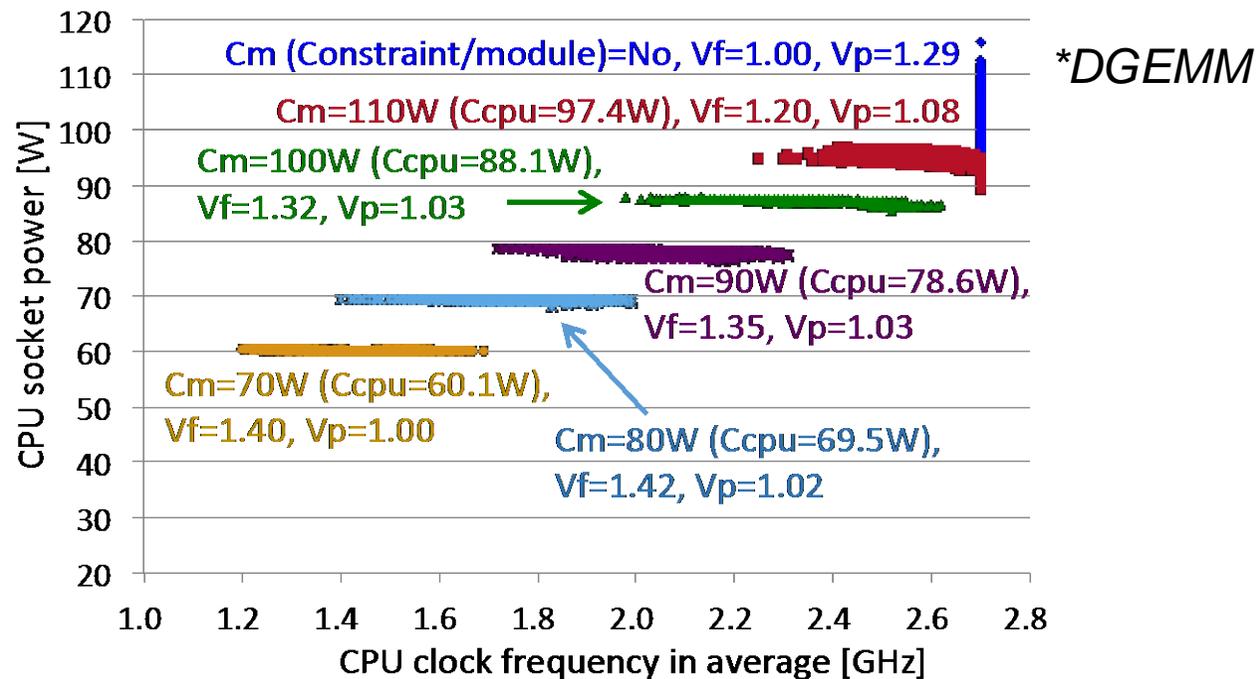
Vp : Max. power variation (Max.-power / Min.-power)

Difference of module-wise power: as large as 30%

# Effect on Performance under Power-Cap (1/2)

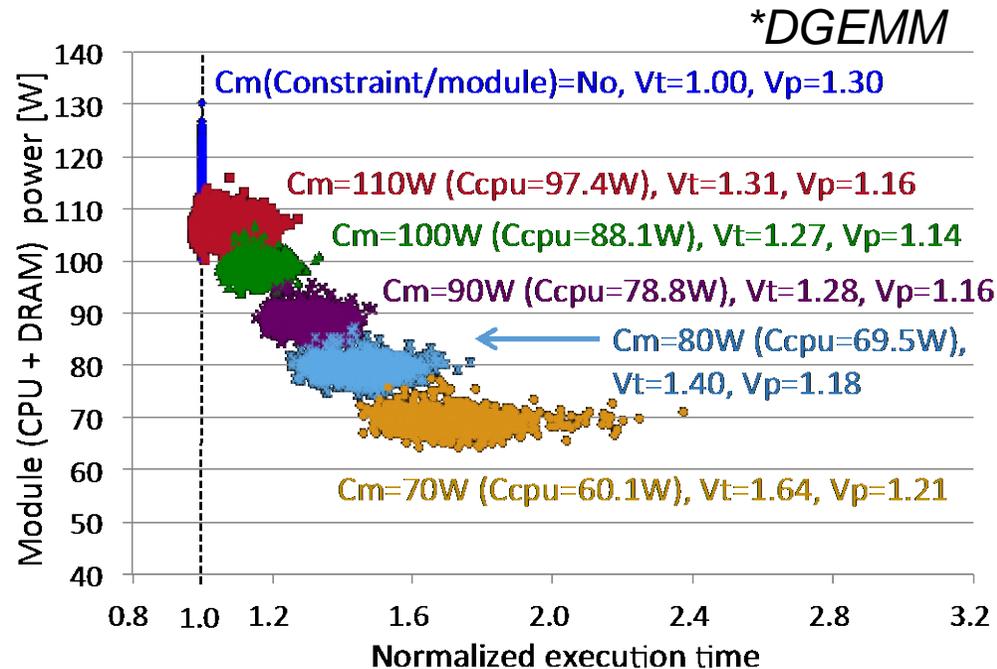
For Over-provisioned systems, power-cap is applied, then..

- Large variation of clock-frequency under the same power-cap
  - CPU power-cap by Intel RAPL interface
  - runtime power is controled by modulating CPU clock-frequency



# Effect on Performance under Power-Cap (2/2)

- Normalized execution time of each CPU-Memory module
- Power variation becomes execution time variation with power-cap



Vt : Max. execution time variation (Max. exec. time / Min exec time)

Difference of module-wise execution time: as large as 40%

# Problem and Our Goal

- Problem

- Manufacturing variability leads to performance degradation for HPC applications
- Performance variation between processors affects well load-balanced parallelized applications



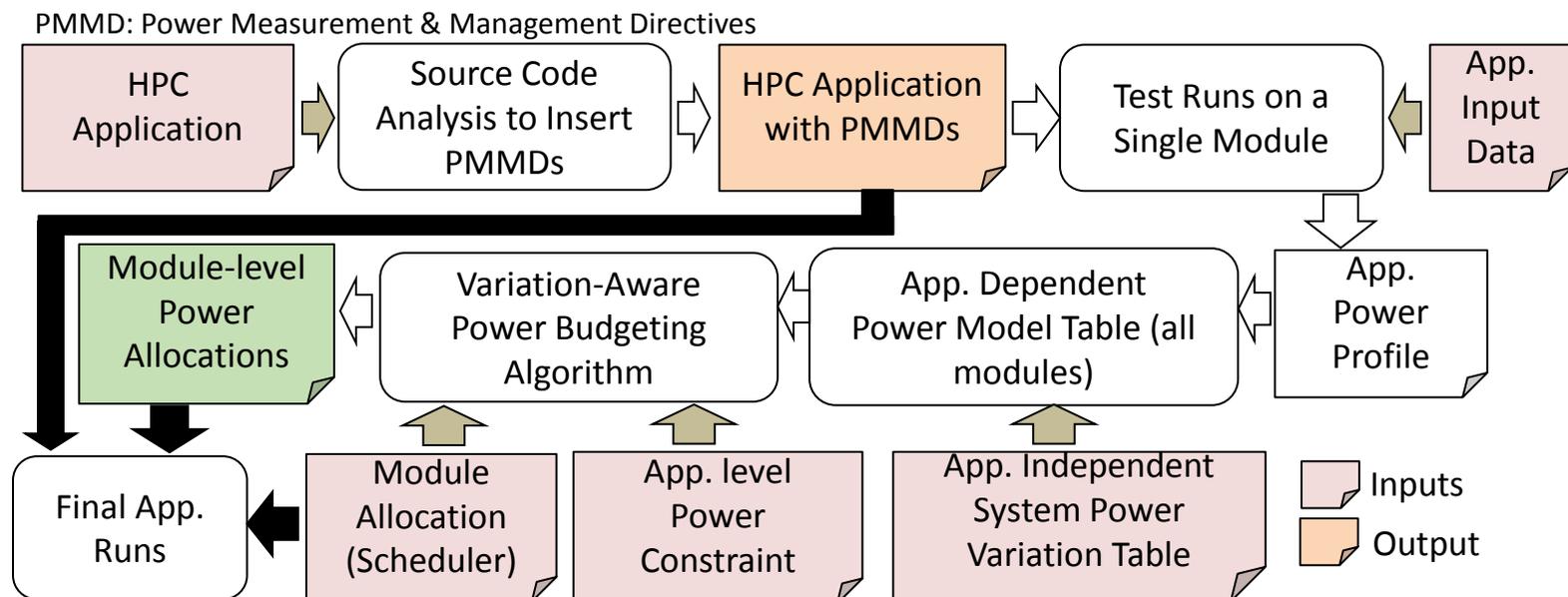
- Our goal

- mitigating the impact of manufacturing variability on performance for HPC apps. under power constraint
- Optimal allocation of power budget for each CPU module taking power variation into account
- Optimal: performance of all the modules are the same

**→ Variation-Aware Power Budgeting**

# Variation Aware Power-Budgeting Strategy

- Required information
  - power-performance relationship of all modules for each application (not easy)
- Basic strategy of the optimization
  1. Pre-characterizing power variation of each module by using micro-benchmarks
  2. Profiling power-performance trend of target application on a single node
  3. Creating app. dependent power-performance model taking variation into account
  4. Calculating power-budget for each node which maximizes performance



# Variation Calibration and Power Model

Created with micro-benchmarks in advance (ex. at system install time)

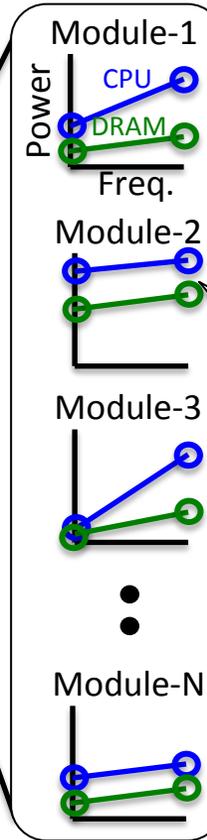
Generated by PVT and profiling results

Application-independent Power Variation Table (PVT)

Module ID	Power on Max. CPU Freq.		Power on Min. CPU Freq.	
	CPU	DRAM	CPU	DRAM
1	0.9	0.8	0.8	0.8
2	1.1	1.1	1.3	1.2
⋮	⋮	⋮	⋮	⋮
<b>k</b>	<b>1.2</b>	<b>1.1</b>	<b>1.4</b>	<b>1.3</b>
⋮	⋮	⋮	⋮	⋮
N	0.8	0.9	1.1	1.2

Application-dependent Power Model Table (PMT)

Module ID	Power on Max. CPU Freq.		Power on Min. CPU Freq.	
	CPU	DRAM	CPU	DRAM
1	90	22	40	12
2	110	30	65	18
⋮	⋮	⋮	⋮	⋮
<b>k</b>	<b>120</b>	<b>30</b>	<b>70</b>	<b>20</b>
⋮	⋮	⋮	⋮	⋮
N	80	24	55	18



Power model for each module with variation into account

Set power-cap of each module to within total power constraint to balance the performance

Average power predicted by module-k

- CPU
- Max. Freq.: 100 W (120 W / 1.2)
- Min. Freq.: 50 W (70 W / 1.4)
- DRAM
- Max. Freq.: 27 W (30 W / 1.1)
- Min. Freq.: 15 W (20 W / 1.3)

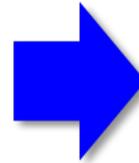
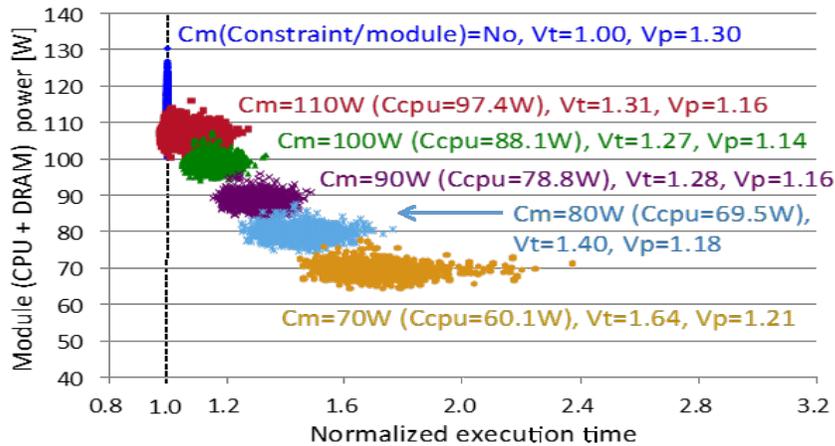
Application-dependent Power on a module (measured)

Module ID	Power on Max. CPU Freq.		Power on Min. CPU Freq.	
	CPU	DRAM	CPU	DRAM
<b>k</b>	<b>120</b>	<b>30</b>	<b>70</b>	<b>20</b>

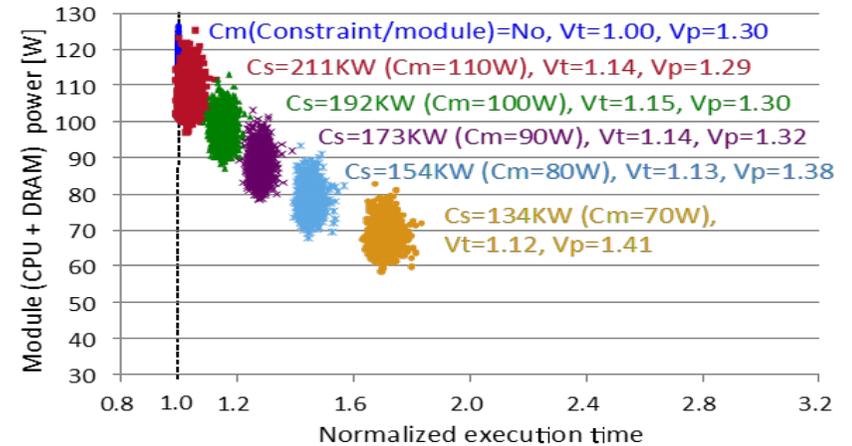
# Effect of Variation-Aware Power Budgeting

- Execution time variation before and after the optimization

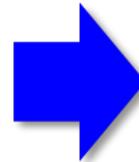
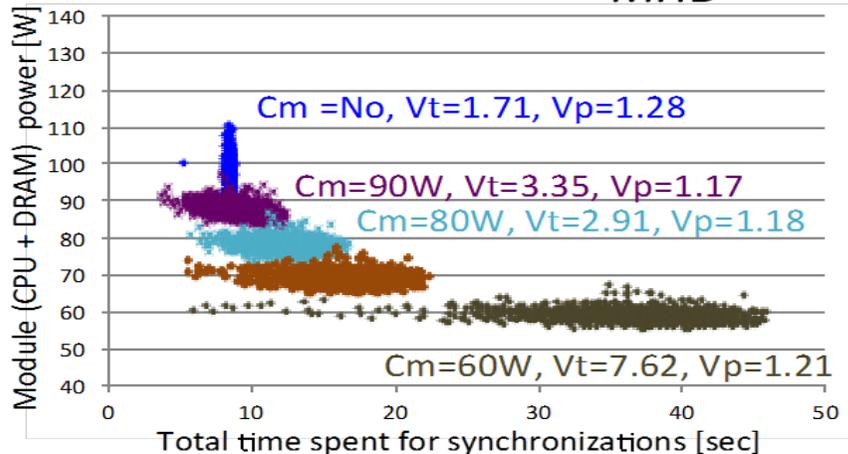
**Before** \*DGEMM



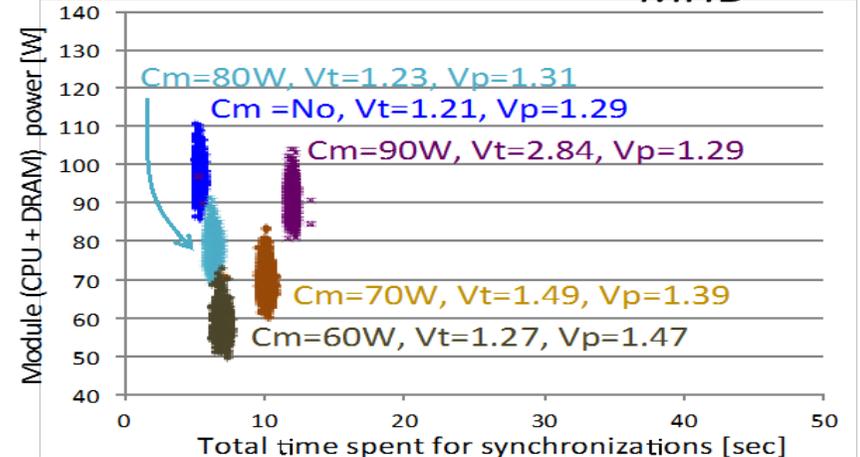
**After** \*DGEMM



**MHD**

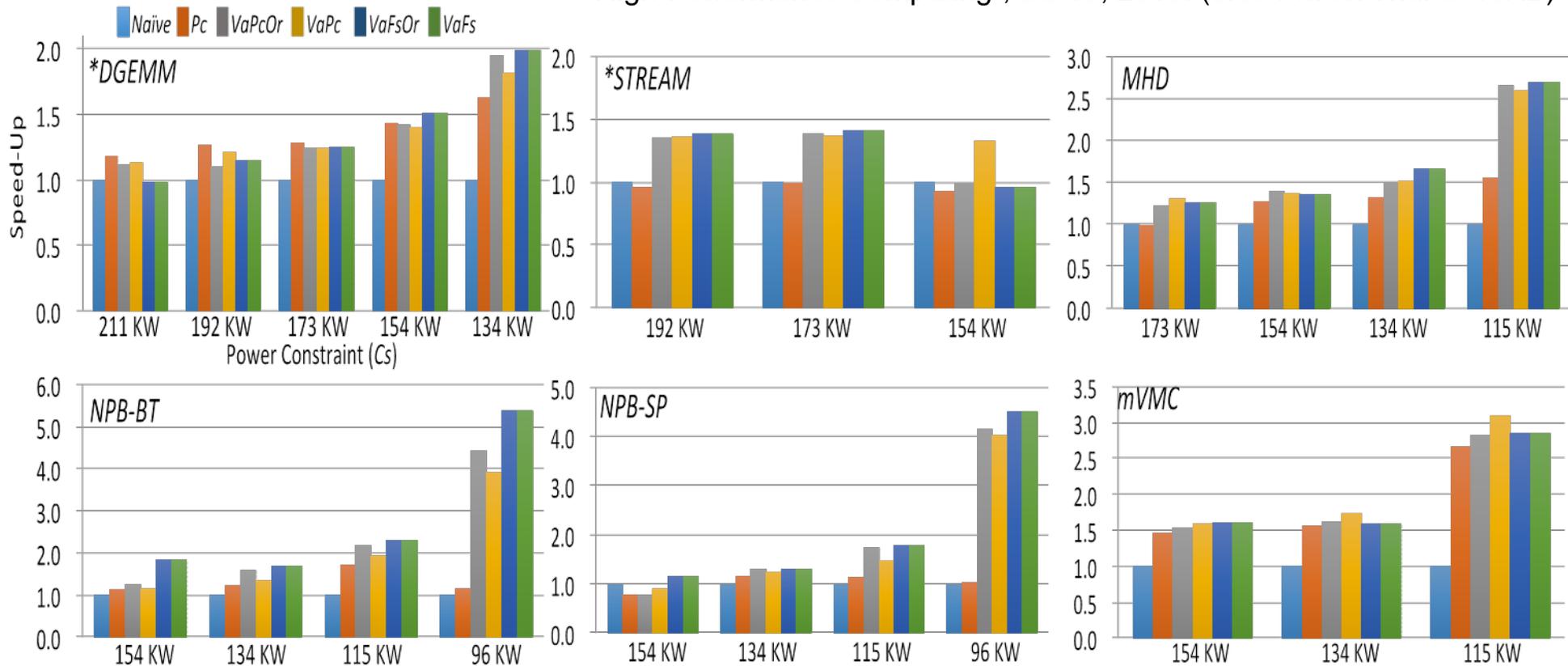


**MHD**



# Effectiveness of Variation Aware Power Budgeting

- Speedup for power budgeting strategies to a naïve method for in various benchmarks Y. Inadomi, et al., “Variation-Aware Power Budgeting in Power-Constrained High-Performance Computing”, SC’15, 2015. (3:30 PM on Thu. at 19AB)



5.4X speedup at maximum, 1.8x on average

# Summary



- Power and energy consumption is a first class design constraint in extremescale systems
  - Power allocation strategy is key to achieve good power and energy efficiency
  - Need international collaboration for HPC eco-system
- Acknowledgement

**Masaaki Kondo**  
The University of Tokyo

**Koji Inoue**  
Kyushu University

**Yuichi Inadomi**  
Kyushu University

**Martin Schulz**  
Lawrence Livermore National Lab.

**Barry Rountree**  
Lawrence Livermore National Lab.

**Tapasya Patki**  
U. of Arizona/LLNL

**David Lowenthal**  
University of Arizona

**Satoshi Matsuoka**  
Tokyo Institute of Technology

**Toshio Endo**  
Tokyo Institute of Technology

**Akira Nukada**  
Tokyo Institute of Technology

**Todd Gamblin**  
Lawrence Livermore National Lab.