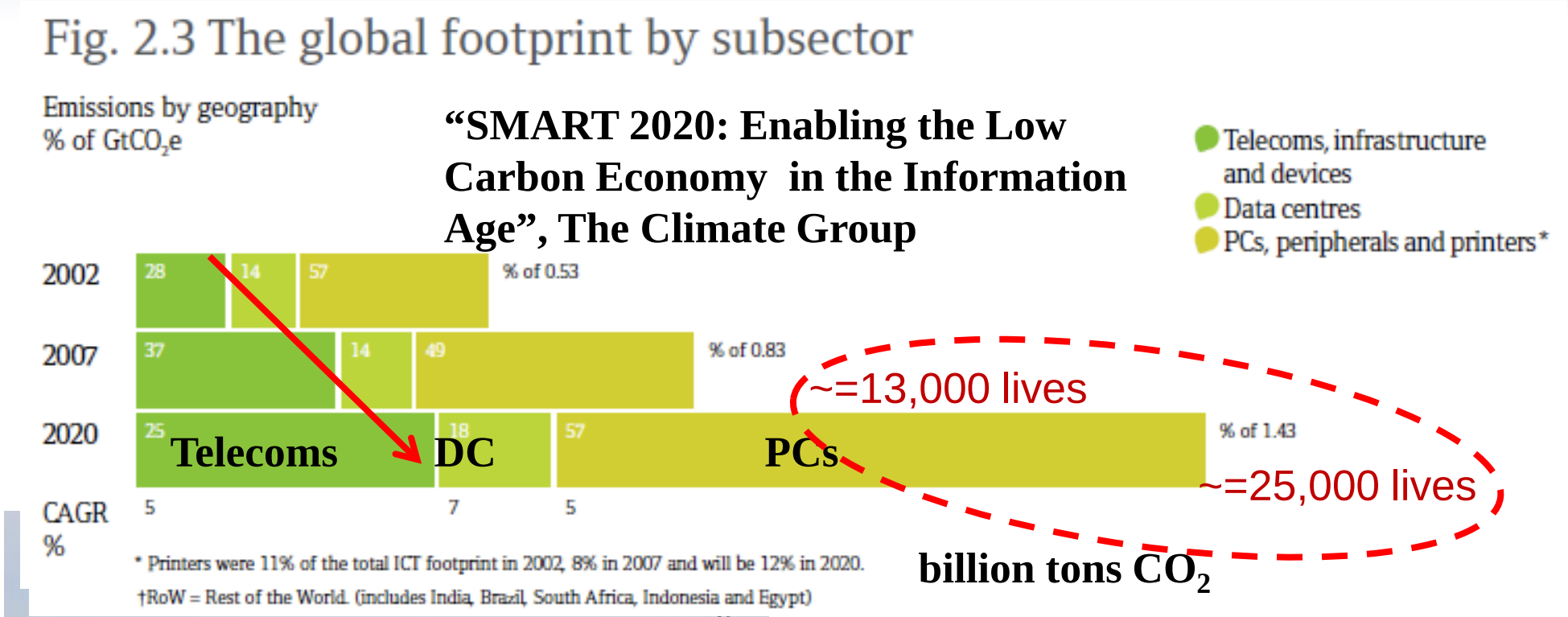
The Japanese Efforts in production Green HPC

Satoshi Matsuoka
Professor
Global Scientific Information and Computing (GSIC) Center
Tokyo Institute of Technology
Fellow, Association for Computing Machinery (ACM)

SC15 EEHPCG Workshop
Austin TX
2015/11/17

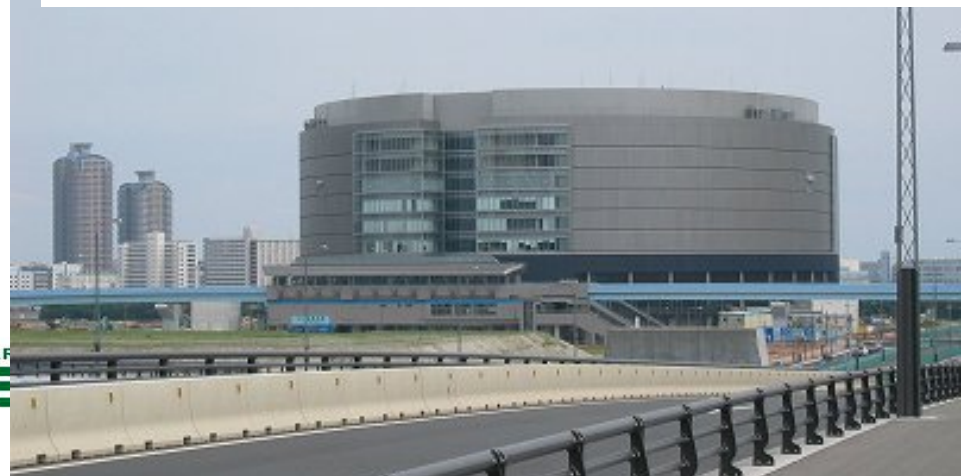
2020 IT Carbon Footprint~=Energy

(Original slide courtesy Horst Simon @ DoE-LBL)



"SMART 2020: Enabling the Low Carbon Economy in the Information Age", The Climate Group

billion tons CO_2



140MW @Tokyo Cloud Datacenter

> 100m diameter

9 stories 140,000 sq. ft.

JST-CREST "MegaScale" [2003-2007]

- *"Embedded tech. meets HPC"*
- Increased importance w/emphasis on mobile/ubiquitous computing
- Attain ~x2 with effective **DVS**

=> 2004 CTWatch J. (Matsuoka ed.)

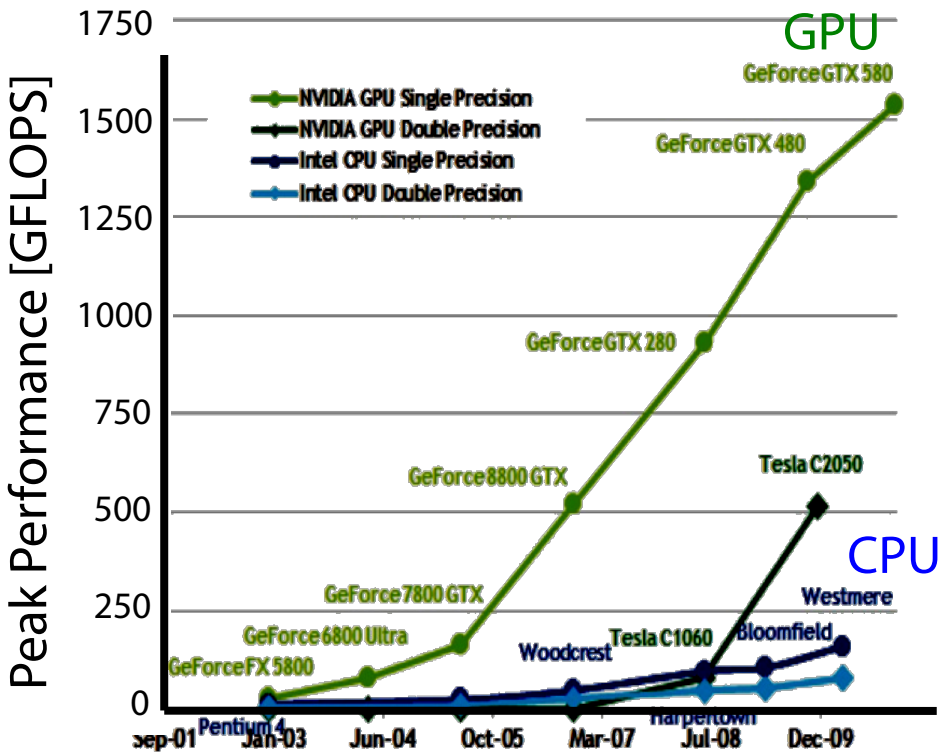
- Can't attain x10 tho…
  - Low hanging fruit achieved with simple DVS
  - Memory, NW, I/O… s tarts to dominate
  - Simple slow cores limited to weak scaling
  - Conclusion: new devices, new architectures, cooling, apps need to be exploited and optimized as a whole

JST-CREST MegaProto (SC2005)
17Sockets Effecion/1U
64GFlops/300W

TSUBAME1.0 June 2006
Asia No.1, World #7
38.18Teraflops(Top500)
1MW power

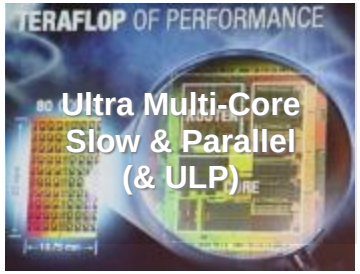
2010 TSUBAME2.0
Petaflop Objective but
Constant Power
→x25 Power Efficiency

CPU vs. GPU Perf/Watt
JST-CREST ULP-HPC Project

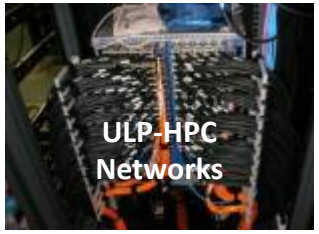


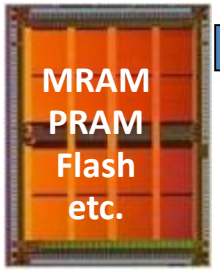
GPU ~x5~6 compute & memory bandwidth advantage
same power
operational test w/TSUBAME1.2

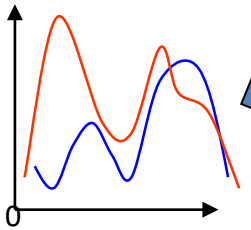
JST-CREST “Ultra Low Power (ULP)-HPC” Project 2007-2012



ULP-HPC SIMD-Vector (GPGPU, etc.)







Low Power
High Perf Model

Power Optimize using Novel Components in HPC

Auto-Tuning for Perf. & Power

モデルと実測の Bayes 的融合

- Bayes モデルと事前分布

$$y_i \sim N(\mu_i, \sigma_i^2)$$

$$\mu_i \mid \beta, \sigma_i^2 \sim N(x_i^T \beta, \sigma_i^2 / \kappa_0)$$

$$\sigma_i^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$

モデルによる所要時間の推定

- n 回実測後の事後予測分布

$$y_i \mid (y_{i1}, y_{i2}, \cdots, y_{in}) \sim t_{\nu_n}(\mu_{in}, \sigma_{in}^2 \kappa_{n+1} / \kappa_n)$$

$$\nu_n = \nu_0 + n, \quad \kappa_n = \kappa_0 + n, \quad \mu_n = (\kappa_0 x_i^T \beta + n\bar{y}_i) / \kappa_n$$

$$\nu_n \sigma_n^2 = \nu_0 \sigma_0^2 + \sum (y_m - \bar{y}_i)^2 + \kappa_0 n (\bar{y}_i - x_i^T \beta)^2 / \kappa_n$$

所要時間の実測データ

ABCLibScript: アルゴリズム選択

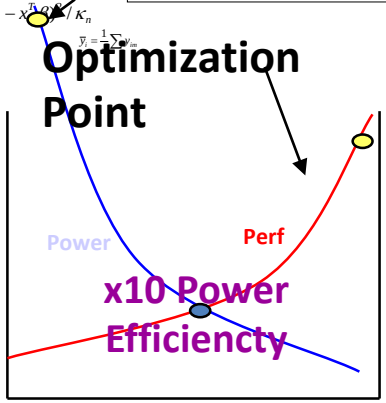
```
!ABCLib$ static select region start
!ABCLib$ parameter (in CacheS, in NB, in NPrc)
!ABCLib$     select sub region start
!ABCLib$     according estimated
!ABCLib$     (2.0d0*CacheS*NB)/(3.0d0*NPrc)
        対象1（アルゴリズム1）
!ABCLib$     select sub region end
!ABCLib$     select sub region start
!ABCLib$     according estimated
!ABCLib$     (4.0d0*ChcheS*dlog(NB))/(2.0d0*NPrc)
        対象2（アルゴリズム2）
!ABCLib$     select sub region end
!ABCLib$ static select region end
```

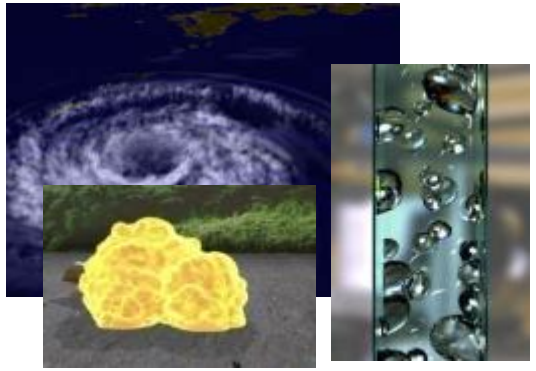
実行起動前自動チューニング指定、アルゴリズム選択処理の指定

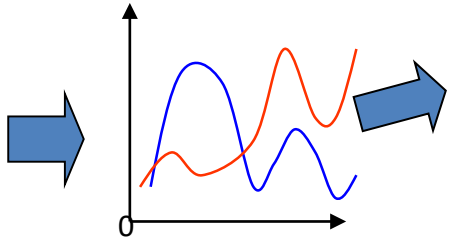
コスト定義関数で使われる入力変数

コスト定義関数

対象領域1、2







Perf. Model
Algorithms

Power-Aware and Optimizable Applications

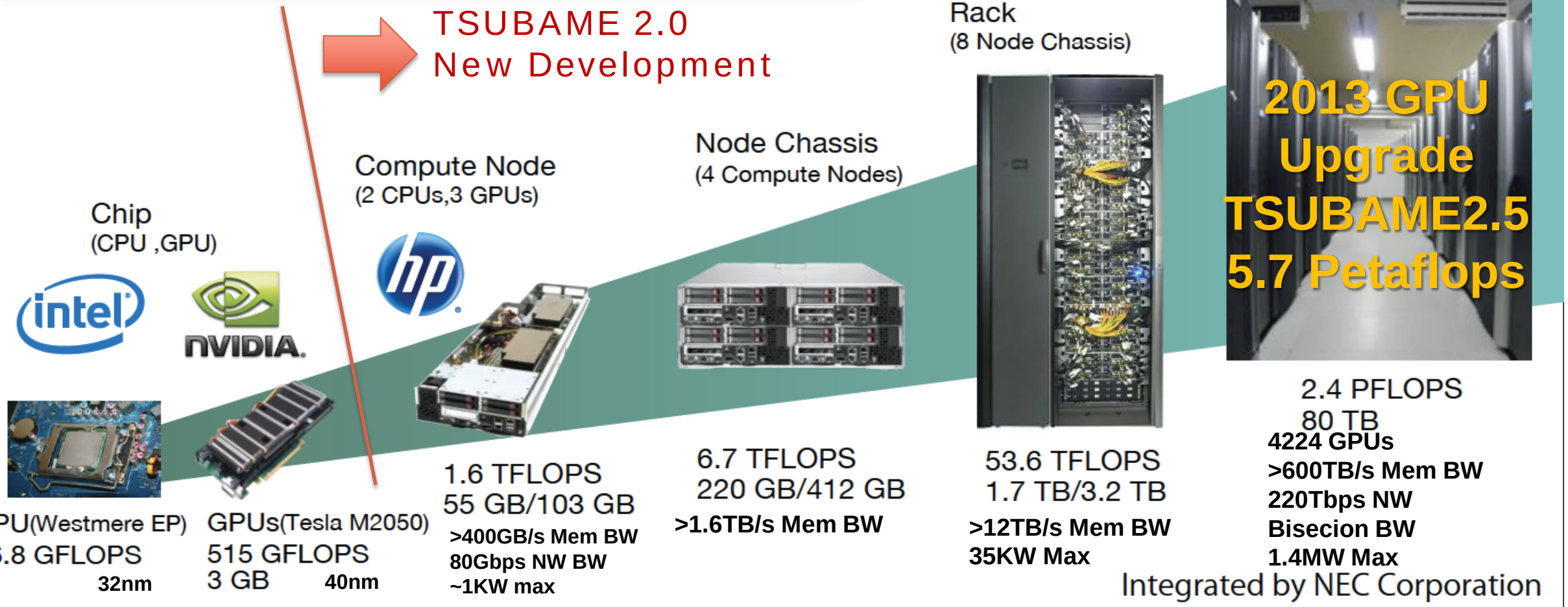
x1000
Improvement in 10 years

Aggressive Power Saving in HPC

Methodologies	Enterprise/Business Clouds	HPC
Server Consolidation	Good	NG!
DVFS (Dynamic Voltage/Frequency Scaling)	Good	Poor
New Devices	Poor (Cost & Continuity)	Good
New Architecture	Poor (Cost & Continuity)	Good
Novel Cooling	Limited (Cost & Continuity)	Good (high thermal density)

TSUBAME2.0 Nov. 1, 2010
"The Greenest Production Supercomputer in the World"

- GPU-centric (> 4000) high performance & low power
- Small footprint (~200m2 or 2000 sq.ft), low TCO
- High bandwidth memory, optical network, SSD storage…



TSUBAME2.0 Very Green…

"Greenest Production Supercomputer in the World"

the Green 500 (#3 overall)

Nov. 2010, June 2011

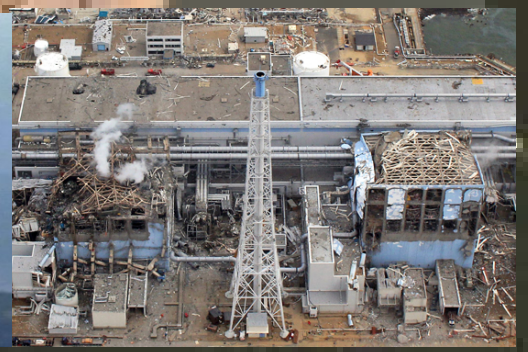
(#4 Top500 Nov. 2010)

<<

3 times more power efficient than a laptop!

It is not just about cost, it is about sustainability!

- WHO estimate: **1.6 million** deaths each year from indoor pollution (Coal&Oil)
  - http://www.who.int/mediacentre/factsheets/fs292/en/index.html
  - 161 deaths/TWh Worldwide
  - Excluding Global Warming effects
- "Clean (not Green) Energy is the solution"
  - Nuclear + Solar + Wind + Hydro …
  - Then 2011 3/11 All hell broke loose in Japan…



3.11 Japan Earthquake

- Top 10 Supercomputers in Japan (**96% perf. Share**) all near Tokyo!
  - 18 of 26 Japanese SCs on the Top500

- All had to be stopped due to rolling blackouts
- Despite all the research, we were not ready for this…

TOP500 Sublist Generator

R_{max} and R_{peak} values are in TFlops. For more details about other fields, check the TOP500 description.

26 entries found.

Rank	Site	System	Cores	R_{max}	R_{peak}
4	GSIC Center, Tokyo Institute of Technology Japan	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows NEC/HP	73278	1192	2287.63
33	Japan Atomic Energy Agency (JAEA) Japan	BX900 Xeon X5570 2.93GHz , Infiniband QDR Fujitsu	17072	191.4	200.08
42	University of Tokyo/Institute for Solid State Physics Japan	SGI Altix ICE 8400EX Xeon X5570 4-core 2.93 GHz, Infiniband SGI	15360	161.8	180.02
54	Japan Agency for Marine - Earth Science and Technology Japan	SX-9/E/1280M160 NEC	1280	122.4	131.07
59	JAXA Japan	Fujitsu FX1, Quadcore SPARC64 VII 2.52 GHz, Infiniband DDR Fujitsu	12032	110.6	121.28
70	Information Technology Center, The University of Tokyo Japan	Hitachi opteron QC 2.3 GHz Myrinet 10G Hitachi	15104	101.74	138.96
74	Institute of Physical and Chemical Res. (RIKEN) Japan	RIKEN Intergrated Cluster of Clusters, Xeon X5570 2.93GHz, Infiniband DDR Fujitsu	9048	97.94	106.04
88	GSIC Center, Tokyo Institute of Technology Japan	Sun Fire x4600/x6250, Opteron 2.4/2.6 GHz, Xeon E5440 2.833 GHz, ClearSpeed CSX600, nVidia GT200; Voltaire Infiniband NEC/Sun	31024	87.01	163.19
96	Center for Computational Sciences, University of Tsukuba Japan	Appro Xtreme-X3 Server - Quad Opteron Quad Core 2.3 GHz, Infiniband Appro International	10368	77.28	95.39
102	National Institute for Environmental Studies Japan	Asterism ID318, Intel Xeon E5530, NVIDIA C2050, Infiniband NSSOL / SGI Japan	144640	74.84	177.12
131	National Institute for Fusion Science (NIFS) Japan	Hitachi SR16000 Model L2/128, Power6 4.7Ghz, Infiniband Hitachi	4096	56.65	77

2016 H2 TSUBAME3.0 Leading Machine Towards Exa & Big Data

1. "Everybody's Supercomputer" - High Performance (~20 Petaflops, 4~5PB/s Mem, ~Pbit/s NW), innovative high cost/performance packaging & design, in mere 180m²…
2. "Extreme Green" – ~10GFlops/W power-efficient architecture, system-wide power control, advanced cooling, future energy reservoir load leveling & energy recovery
3. "Big Data Convergence" – Extreme high BW &capacity, deep memory hierarchy, extreme I/O acceleration, Big Data SW Stack for machine learning, graph processing, …
4. "Cloud SC" – dynamic deployment, container-based node co-location & dynamic configuration, resource elasticity, assimilation of public clouds…
5. "Transparency" - full monitoring & user visibility of machine & job state, accountability via reproducibility

2006 TSUBAME1.0
80 Teraflops, #1 Asia #7 World
"Everybody's Supercomputer"

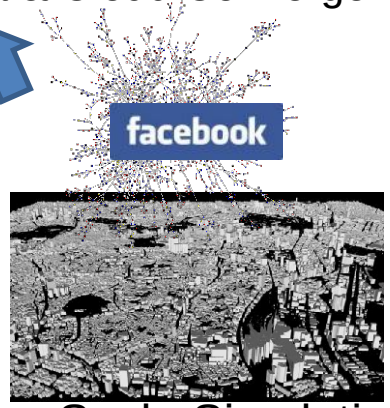
2010 TSUBAME2.0
2.4 Petaflops #4 World
"Greenest Production SC"

2011 ACM Gordon Bell Prize

2013
TSUBAME2.5
upgrade
5.7PF DFP
/17.1PF SFP
20% power
reduction

2013 TSUBAME-KFC
#1 Green 500

2016 TSUBAME3.0
~20PF(DFP) 4~5PB/s Mem BW
10GFlops/W power efficiency
Big Data & Cloud Convergence



Large Scale Simulation
Big Data Analytics
Industrial Apps

2020 Global Target

1 Exaflop, 200TeraByte/s@ 20MW

= 50 GigaFlops/W, 10GB/s / W

x1000~x2500 improvement!

Machineg	Power (incl. cooling)	Linpack Perf (PF)	Linpack MFLOPs/W	Factor	Total Mem BW TB/s (STREAM)	Mem BW MByte/S / W	Factor
Earth Simulator 1	10MW	0.036	3.6	13,400	160	16	312
Tsubame1.0 (2006Q2)	1.8MW	0.038	21	2,368	13	7.2	692
Tsubame2.0 (2010Q4)	**1.8MW**	**1.2**	**667**	**75**	**440**	**244**	**20**
			x32			***x34***	
K Computer (2011Q2)	~16MW	10	625	80	3300	206	24
BlueGene/Q (2012Q1)	~12MW?	17	~1400	~35	3000	250	20
TSUBAME2.5 (2013Q3)	**1.4MW**	**~2.8**	**~2000**	**~25**	**802**	**572**	**8.7**
TSUBAME3 (2016Q2)	**1.8MW**	**~18**	**~10,000**	**~5**	**~50K**	**~2800**	**~4**
			~x75			***~x40***	
EXA (2020)	20MW	1000	50,000	1	200K	10,000	1

TSUBAME-KFC/DL(update): TSUBAME3 Prototype [ICPADS2014]

Oil Immersive Cooling＋ Hot Water Cooling + High Density Packaging + Fine-Grained Power Monitoring and Control¥

High Temperature Cooling

Oil Loop 35~45℃

⇒ Water Loop 25~35℃

(c.f. TSUBAME2: 7~17℃)

Cooling Tower:

Water 25~35℃

⇒ To Ambient Air

168 NVIDIA K80, one rack

High Density Oil Immersion

413.5TFlops (DFP)

~1.5 Petaflops (SFP)

Greenest on the Top500

Experimental Container Facility

20 feet container (16m^2)

Fully Unmanned Operation

TSUBAME-KFC #1 Green 500 List (Nov. 2013)

- 1st achievement as Japanese supercomputer
- #1 again in June 2014
- 10GFLOPS/W incl. Cooling for TSUBAME3, + Energy Recovery

The Green500 List

Listed below are the November 2013 The Green500's energy-efficient supercomputers ranked from 1 to 10.

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	4,503.17	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	27.78
2	3,631.86	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
3	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray 3623G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
4	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Level 3 measurement data available	1,753.66
5	3,130.95	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x	81.41
6	3,068.71	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.930GHz, Infiniband QDR, NVIDIA K20x	922.54
7	2,702.16	University of Arizona	iDataPlex DX360M4, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR14, NVIDIA K20x	53.62
8	2,629.10	Max-Planck-Gesellschaft MPI/IPP	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	269.94
9	2,629.10	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	55.62
10	2,358.69	CSIRO	CSIRO GPU Cluster - Nitro G16 3GPU, Xeon E5-2650 8C 2.000GHz, Infiniband FDR, Nvidia K20m	71.01

TSUBAME-KFC

TSUBAME 2.5

* Performance data obtained from publicly available sources including TOP500

Ranks 1 - 100

Ranks 101 - 200

Ranks 201 - 300

Ranks 301 - 400

Ranks 401 - 500

Downloads

Microsoft Excel™ Spreadsheet (.xls)

Comma Separated Values (.csv)

THE GREEN 500™
sponsored by SUPERMICRO

This certificate is in recognition of your organization's achievements in reducing the environmental impact of high-performance computing.

GSIC Center, Tokyo Institute of Technology

is ranked

1st

on the world's Green500 List of computer systems as of

November 2013

Wu-chun Feng, Co-Chair

Kirk Cameron, Co-Chair

Green500 Measurement Changing in 2016

- New Level1
  - Measures the whole Linpack run, not parts
  - Include interconnect and switches
  - Other changes

- New measurement more accurate account of power efficiency
  - FLOPS/W = FLOP/J if constant

- TSUBAME-KFC/DL measurements: about 6% difference between old and new L1

How do we sustain x1000?

Process Shrink x100

X

Many-Core GPU Usage x5

X

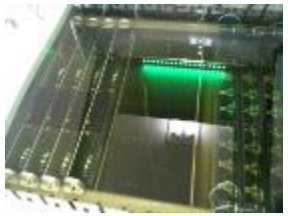
DVFS & Other LP SW x1.5

X

Efficient Cooling x1.4

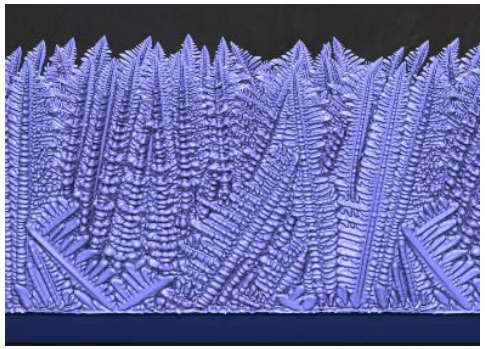
x1000 !!!

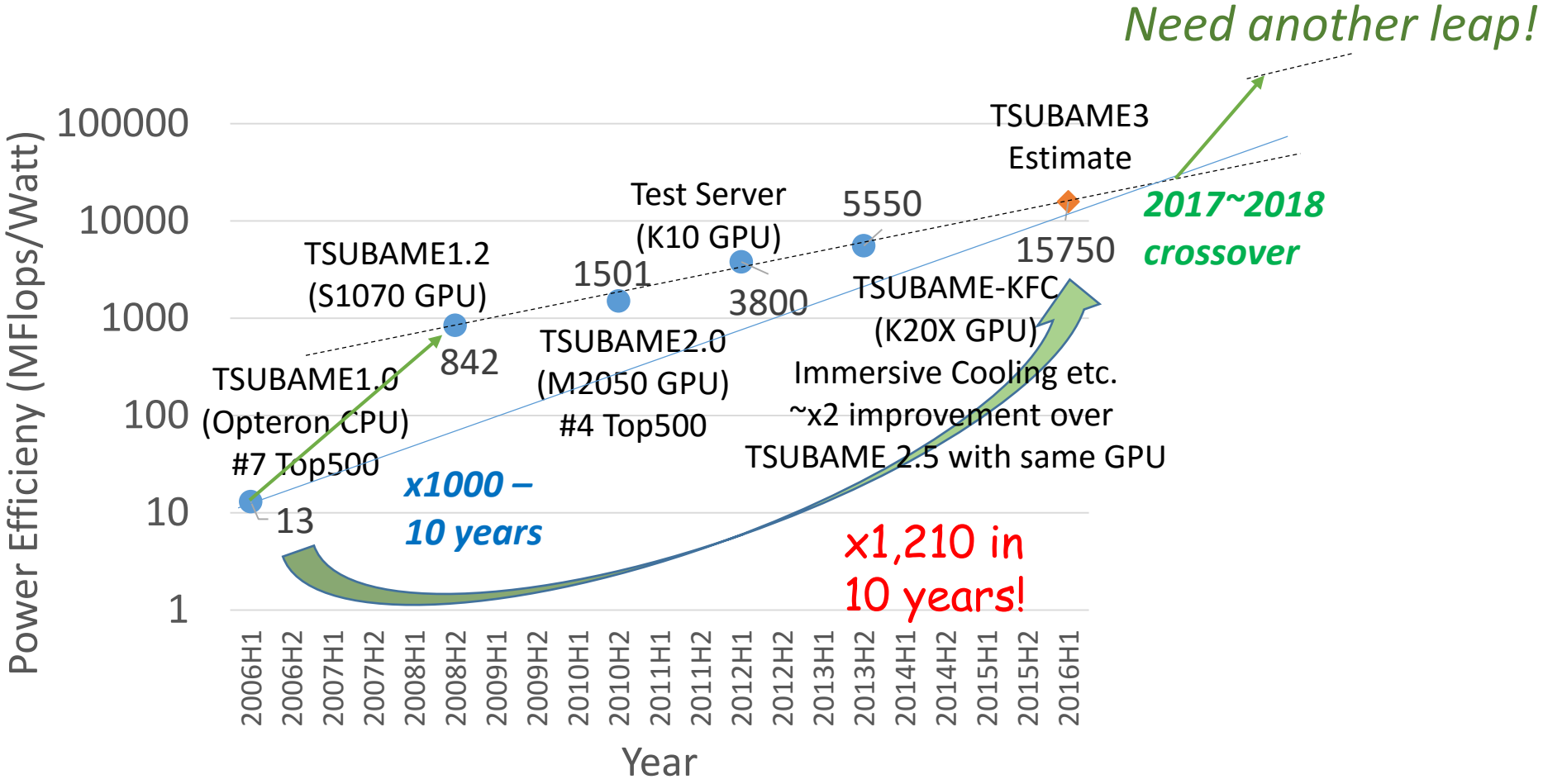
ULP-HPC Project 2007-12



Ultra Green Supercomputing Project 2011-15

Power Efficiency of GB Dendrite Simulation since 2006
Good news and bad news towards Post-Moore

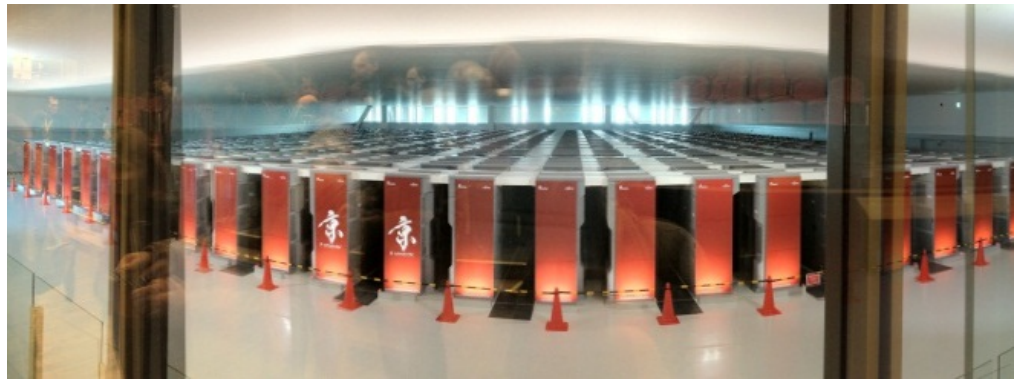




Measured for the 2011 Gordon Bell Award Dendritic Solidification App
Flop/s/W = Total #Flops / J = energy to solution given same problem

TSUBAME4 beyond 2021~2022 K-in-a-Box (Golden Box) Towards Post-Moore Performance Growth

1/1000 Size, 1/150 Power, 1/500 Cost, x5 DRAM+ NVM Memory



10 Petaflops, 10 Petabyte Hiearchical Memory (K: 1.5PB), 10K nodes

50GB/s Interconnect (200-300Tbps Bisection BW)

(Conceptually similar to HP “The Machine”)

Datacenter in a Box

Large Datacenter will become “Jurassic”