

To Fear or Not to Fear Exascale

Satoshi Matsuoka

Global Scientific Information and Computing
Center (GSIC)

Tokyo Institute of Technology (Tokyo Tech.)

CCGSC 2010 Panel

8/Sep/2010

How can we overcome the fear?

- Which fears are mistaken (after all, many were convinced that petascale systems would be impossible without new programming models)?
- Conversely, which problems apply at a smaller scale, and hence can be addressed now and provide near-term benefits?
- Which problems are (nearly) unique to Exascale? How do we build/test/improve algorithms, software, and applications? For example, do we need to build a much more sophisticated simulation environment?

How can we build real excitement?

- How do we provide evidence that Exascale systems will work with applications?
- How do we demonstrate that Exascale systems can enable new application areas (after all, Exascale systems may be greatly different in architecture - will that be a virtue)?
- In all of the above, how do we move past qualitative statements to quantitative predictions?

Which fears are mistaken?

- Most fears are likely real
- Some fears may not be...?

TSUBAME 2.0 Compute Nodes Details

“Thin” nodes implement the “vector-scalar” hybrid high-bandwidth design combining NVIDIA “Fermi” Tesla GPUs + Intel Westmere CPU in a new, customly architected & designed high bandwidth nodes for Multi-GPU computing

Highly dense, efficient power & thermals, extremely reliable, extensive monitoring & mgmt

Thin Compute Nodes

IB QDR x2



NVIDIA M2050 (Fermi)
515GFLOPS/GPU
3GPU/node

Custom Designed with Hewlett Packard
CPU: Intel Westmere-EP 2.93GHz x2 (12cores/node) TB :3.196GHz
Memory: 55.8GB(=52GiB) DDR3 1333MHz
103GB(=96GiB) DDR3 1333MHz
SSD : 60GB x2 (120GB/node) ✖Memory 55.8GB nodes
120GB x2 (240GB/node) ✖Memory 103GB nodes

1408nodes : 215.99TFlops ✖Turbo boost

4224GPUs : 2175.36TFlops

Total: 2391.35TFLOPS

Memory: 80.55TB

SSD: 173.88TB

Medium Compute Nodes

IB QDR



PCI-e Gen2x16 x2
✖for NVIDIA Tesla
S1070 GPUs

HP 4-Socket Server
CPU: Intel Nehalem-EX 2.0GHz x4 32core/node
Memory: 137GB(=128GiB) DDR3 1066MHz
SSD : 120GB x4 (480GB/node)

24nodes:6.14TFlops

Total: 6.14TFlops

Fat Compute Nodes

IB QDR



PCI-e Gen2x16 x2
✖for NVIDIA Tesla
S1070 GPUs

HP 4-Socket Server
CPU: Intel Nehalem-EX 2.0GHz x4 (32core/node)
Memory: 274GB(=256GiB) DDR3 1066MHz
549GB(=512GiB) DDR3 1066MHz
SSD : 120GB x4 (480GB/node)

10nodes:2.56TFlops

Total: 2.56TFlops

CPU : 224.69TFlops
GPU : 2175.36TFlops

Total
2.4PFlops
(200TB SSD)

**Do not make photo
public until
October 2010**



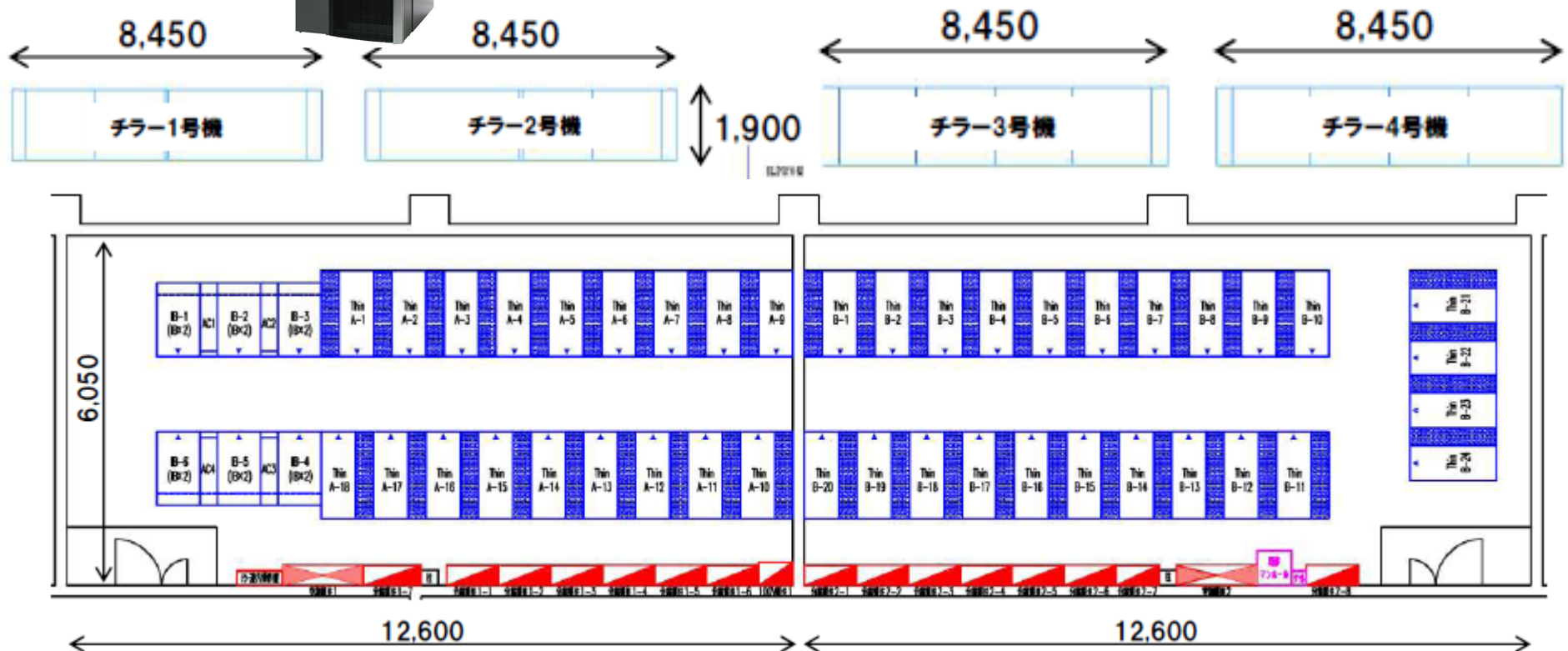
Fear of Space?

TSUBAME2.0 (Circa 2010) Layout (Less than 200m² for main compute nodes)



~= **Entire Earth Simulator**
(rack = 50TF)

BG/Q will even be denser

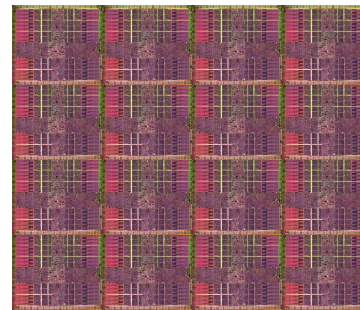
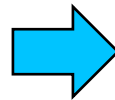
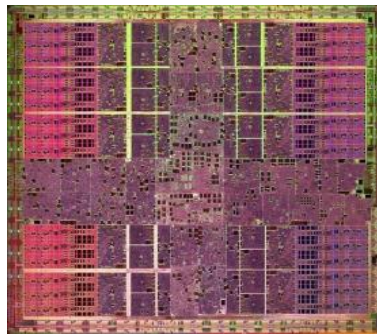




Not to say that Industry Standards are not viable: scaling TSUBAME2.0 to Exaflop

- TSUBAME2.0: 32-40nm, 2.4PF, ~50 racks, 1.5MW = x400 scaling?
- x25 physical scaling now (1000 racks, 40MW)
 - >3000-4000m², 1000 tons
- x16 semiconductor feature scaling 2016~2017

2008	2009	2010	2011	2012	2014	2016-17
45nm	40nm	32nm	28nm	22nm	15nm	11nm



Other innovations such as 3-D memory / flash packaging, optical chip-chip connect, multi-rail optical interconnect etc.

45nm

11nm, x16 transistors & FLOPs

But what about the network? 3-40,000 nodes?

Japanese 10 PF Facility @ Kobe, Japan

**Construction: started in March, 2008 and will complete in May, 2010,
Machine operation late 2011 ~ early 2012**

Computer Wing

Total Floor Area: 17,500m²

2 Computer rooms: 12,600m²

4 Floors (1 underground floor)

=> x4 ES, 2000~4000 racks



Other Facilities

Co-generation System

Water chiller system

Electric Subsystem

**Initially 30MW
capability@2011**

The Technical Challenges

- A. There are three aspects relating to scale**
 - A. Challenges brought on by billion-way parallelism**
E.g., Programming, Algorithms and Amdahl's law, Deepening Hierarchy and Heterogeneity
 - B. Challenges brought on by extra x10 scaling requirements on top of the Moore's law**
E.g. Power Management, Resilience, I/O, ...
 - C. Strong Scaling Challenges:** Challenges brought on by decreasing relative bandwidth, constant including signaling as well as the n^2 vs. n problem
However, parallelism helps---See next slide



Fear of Not Strong Scaling

- Shorten latency as much as possible
 - Extreme multi-core incl. vectors
 - "Fat" nodes, exploit short-distance interconnection
 - Direct cross-node DMA (e.g., put/get for PGAS)
- Hide latency if cannot be shortened
 - Dynamic multithreading (Old: dataflow, New: GPUs)
 - Trade Bandwidth for Latency (so we do need BW...)
 - Departure from simple mesh system scaling
- Change Latency-Starved Algorithms
 - From implicit Methods to direct/hybrid methods
 - Structural locality, extrapolation, stochastics (MC)
 - Need good programming model/lang/system for this...

DOE SC Applications Overview

(following slides courtesy John Shalf @ LBL NERSC)

NAME	Discipline	Problem/Method	Structure
MADCAP	Cosmology	CMB Analysis	Dense Matrix
FVCAM	Climate Modeling	AGCM	3D Grid
CACTUS	Astrophysics	General Relativity	3D Grid
LBMHD	Plasma Physics	MHD	2D/3D Lattice
GTC	Magnetic Fusion	Vlasov-Poisson	Particle in Cell
PARATEC	Material Science	DFT	Fourier/Grid
SuperLU	Multi-Discipline	LU Factorization	Sparse Matrix
PMEMD	Life Sciences	Molecular Dynamics	Particle

Latency Bound vs. Bandwidth Bound?

- How large does a message have to be in order to saturate a dedicated circuit on the interconnect?
 - $N^{1/2}$ from the early days of vector computing
 - Bandwidth Delay Product in TCP

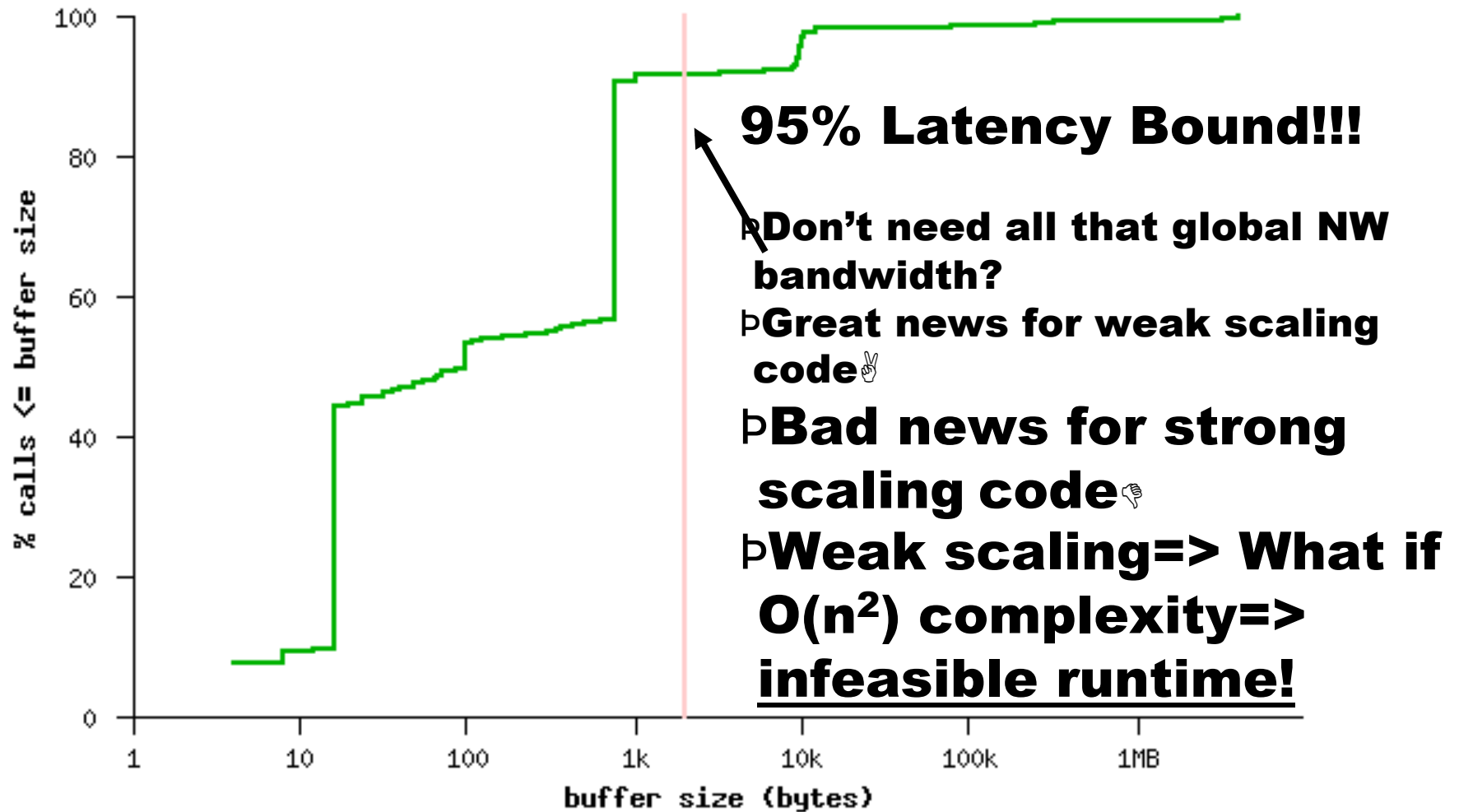
System	Technology	MPI Latency	Peak Bandwidth	Bandwidth Delay Product
SGI Altix	Numalink-4	1.1us	1.9GB/s	2KB
Cray X1	Cray Custom	7.3us	6.3GB/s	46KB
NEC ES	NEC Custom	5.6us	1.5GB/s	8.4KB
Myrinet Cluster	Myrinet 2000	5.7us	500MB/s	2.8KB
Cray XD1	RapidArray/IB4x	1.7us	2GB/s	3.4KB

- Bandwidth Bound if msg size $>$ Bandwidth*Delay
- Latency Bound if msg size $<$ Bandwidth*Delay
 - Except if pipelined (*unlikely with MPI due to overhead*)
- W/HW DMA a few 100ns but not much more

Collective Buffer Sizes are Small(!)

- demise of message passing in strong scaling -

Collective Buffer Sizes for All Codes



(Original slide courtesy John Shalf @ LBL)

Highlights of TSUBAME 2.0 Design (Oct. 2010) w/NEC-HP

- **2.4 PF Next gen multi-core x86 + next gen GPGPU**
 - ▶ 1432 nodes, Intel Westmere/Nehalem EX
 - ▶ 4224 NVIDIA Tesla (Fermi) M2050 GPUs
 - ▶ ~100,000 total CPU and GPU "cores", High Bandwidth
 - ▶ **1.9 million "CUDA cores", 32K x 4K = 130 million CUDA threads(!)**
- **0.72 Petabyte/s aggregate mem BW,**
 - ▶ Effective 0.3-0.5 Bytes/Flop, restrained memory capacity (100TB)
- **Optical Dual-Rail IB-QDR BW, full bisection BW(Fat Tree)**
 - ▶ **200Tbits/s**, Likely fastest in the world, still scalable
- **Flash/node, ~200TB (1PB in future), 660GB/s I/O BW**
 - ▶ >7 PB IB attached HDDs, 15PB Total HFS incl. LTO tape
- **Low power & efficient cooling, comparable to TSUBAME 1.0 (~1MW); PUE = 1.28 (60% better c.f. TSUBAME1)**
- **Virtualization and Dynamic Provisioning of Windows HPC + Linux, job migration, etc.**

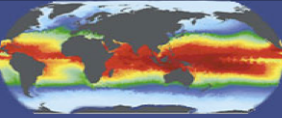


**(From US DoE Exascale PPT by Rick Stevens@ANL)
 Uncertainty quantification further helps in utilizing
 parallelism and enabling scaling.**



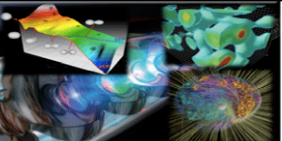
Response surface
Posterior exploration
Finding least favorable priors
Bounds on functionals

Adjoint enabled forward models
Data extraction from model
Local approximations, filtering
Stochastic error estimation



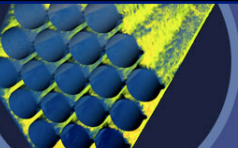
Challenges in Climate Change Science and the Role of Computing at the Extreme Scale
 November 6-7, 2008 · Washington D.C.

“We need to be able to make quantitative statements about the predictability of regional climatic variables that are of use to society.”



Forefront Questions in Nuclear Science and the Role of High Performance Computing
 January 26-28, 2009 · Washington D.C.

“computational techniques and needs complement the scientific areas that will be pursued with extreme scale computing. Examples include ... verification and validation issues for extreme scale computations ”



Science Based Nuclear Energy Systems Enabled by Advanced Modeling and Simulation at the Extreme Scale
 May 11 and May 12, 2009 - Washington DC

“scientists must create new suites of application codes, Integrated Performance and Safety Codes (IPSCs) that incorporate ...integrated uncertainty quantification..”

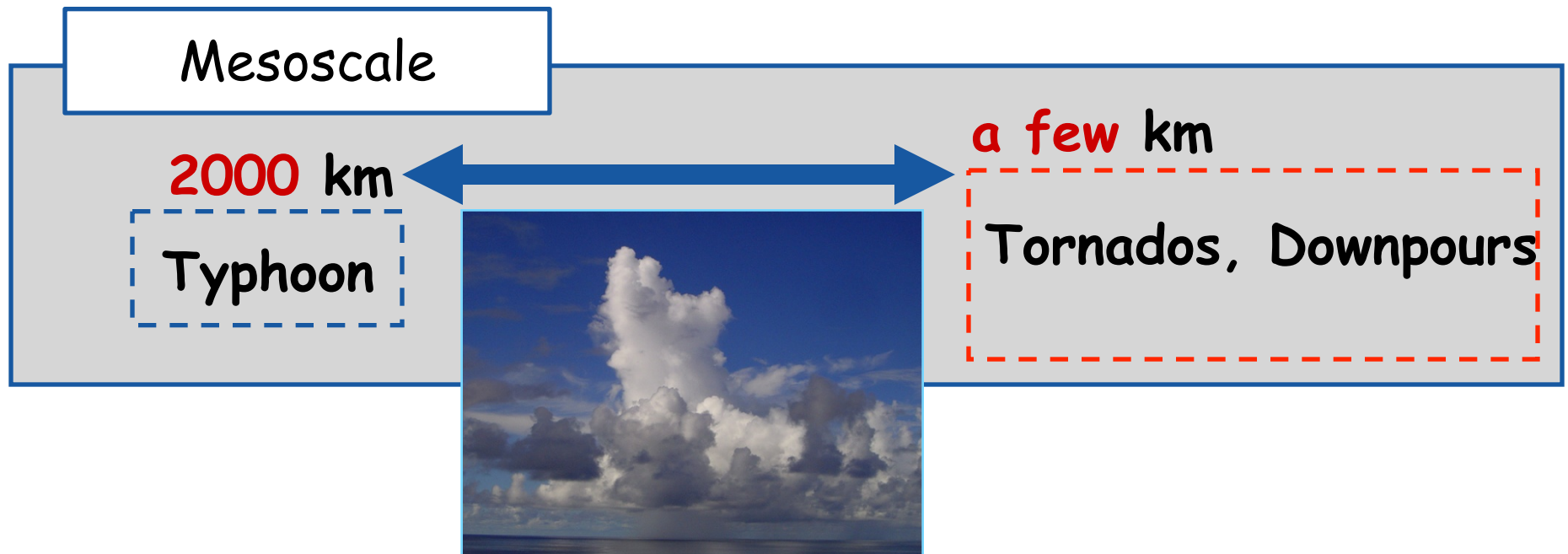


Next Gen Weather Forecast

Mesoscale Atmospheric Model:

Cloud Resolution: 3-D non-static

Compressible equation taking consideration of sound waves.



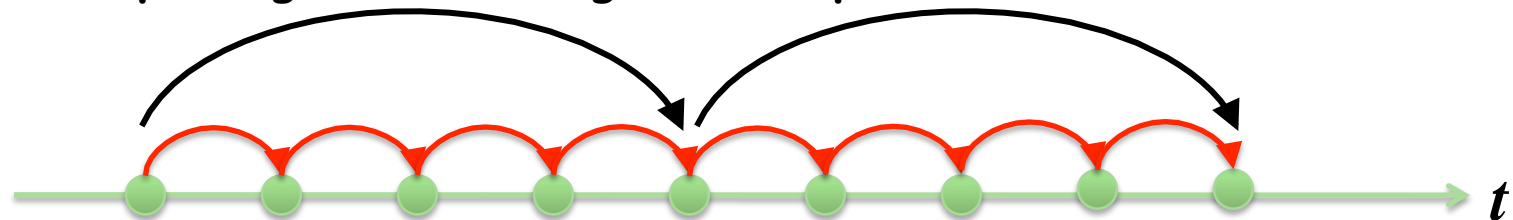
GPU enabling ASUCA [SC10]

- **ASUCA : Next Generation Production Weather Forecast Code (by Japan's National Meteorological Agency)**

Mesoscale production code for real weather forecast

Very similar to NCAR's WRF

Time-splitting method: long time step for flow

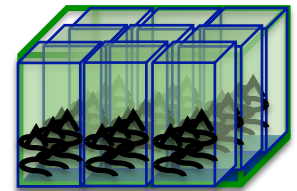


$u, v (\sim 100 \text{ m/s}), w (\sim 10 \text{ m/s}) \ll \text{sound velocity } (\sim 300 \text{ m/s})$

HEVI (Horizontally explicit Vertical implicit) scheme

Horizontal resolution $\sim 1 \text{ km}$

Vertical resolution $\sim 100 \text{ m}$



1-D Helmholtz equation (like Poisson eq.)

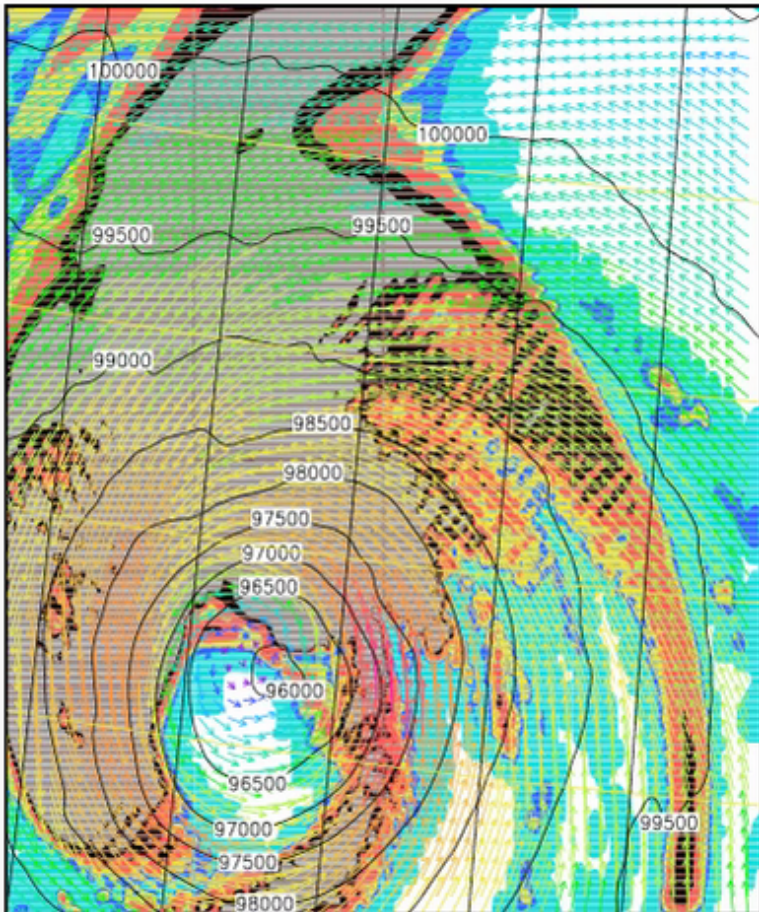
sequential process

*Entire "Core" of ASUCA now ported to GPU (~30,000 lines)
By Prof. Aoki Takayuki's team at Tokyo Tech.*

ASUCA Typhoon Simulation

2km mesh 3164×3028×48

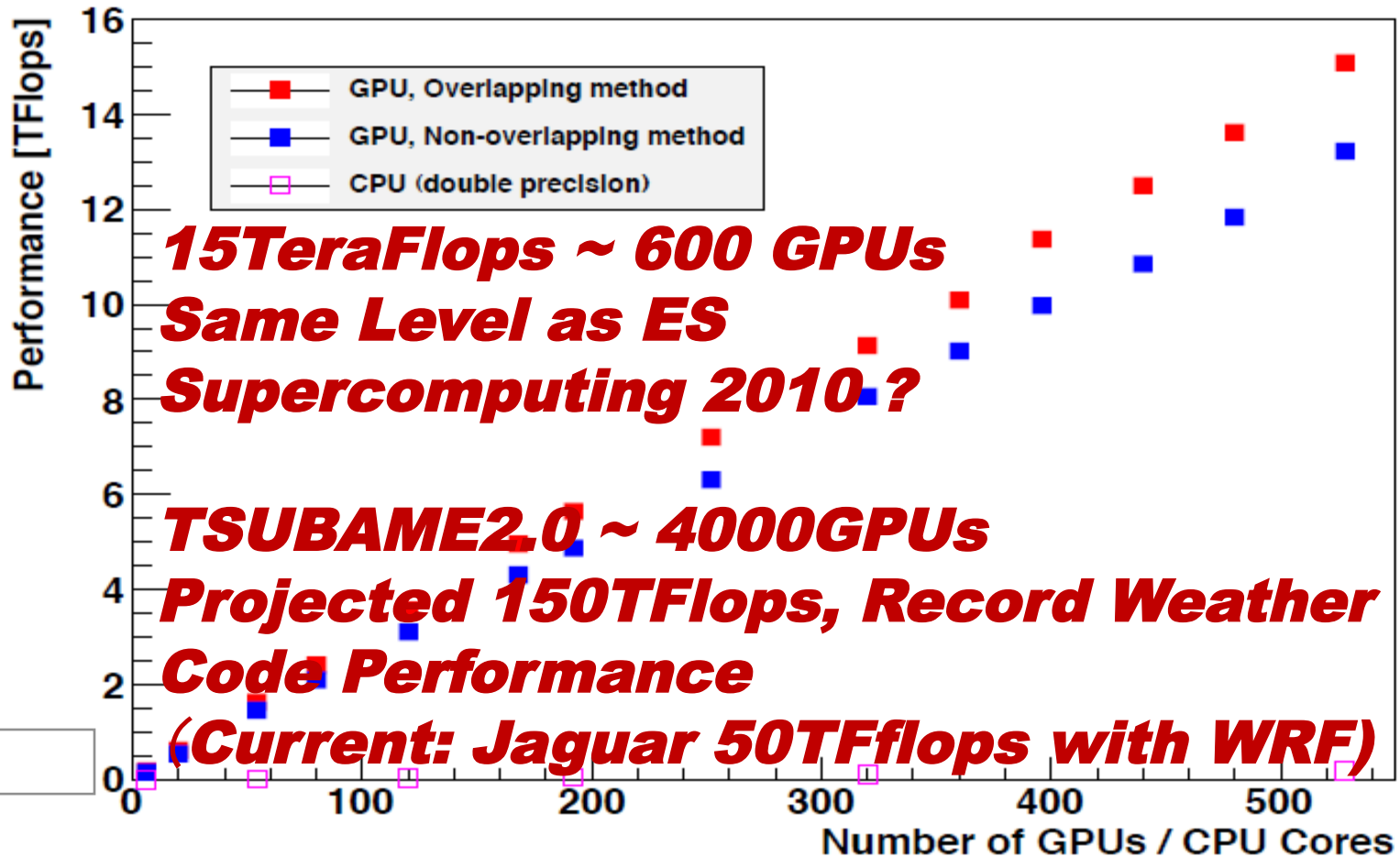
uv and smqr T=1



70 minutes wallclock time
for 6 hour simulation time
(x5 faster)

Expect real-time with
0.5km mesh (for Clouds)
on TSUBAME2.0

ASUCA Multi GPU Performance (TSUBAME1.2)



TSUBAME2 Peak 2.4PF ~ = Jaguar Peak 2.33PF

TSUBAME2 Power ~1.6MW x4 = Jaguar Power ~7MW

Fear of Cost and (Lack of) Excitement?

1. Multi-billion dollar exascale initiative is necessary to develop exascale by 2018. Will there be apps to justify such an initiative?

A. There have been bigger projects before and now Societal costs worth guarantee sustainability of humans

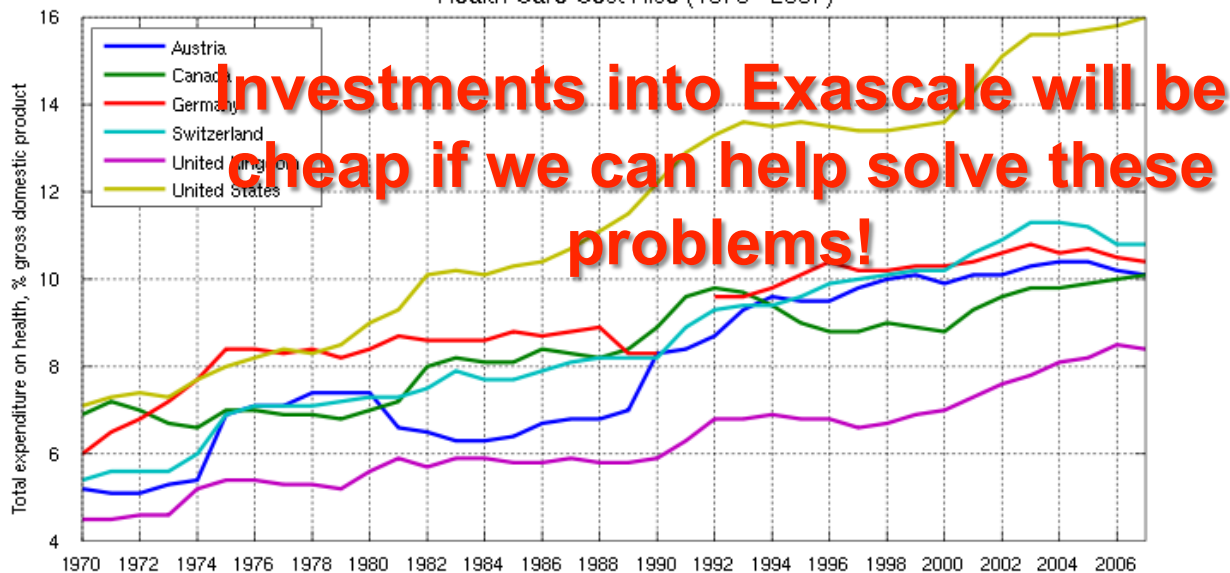
Ex. 1. Space programs; ETA and LHC multi-billion dollar

Ex. 2. US Health Care cost 16% of GDP in 2007

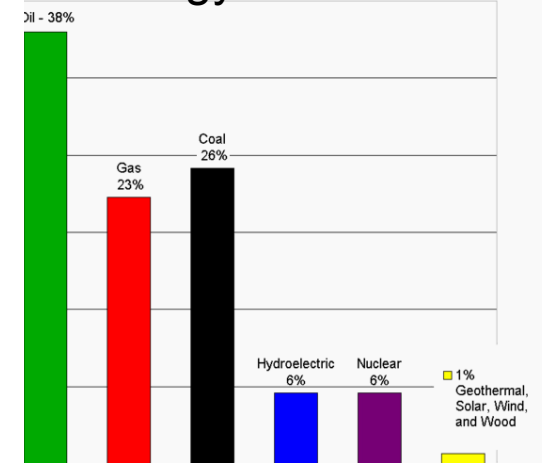
Ex. 3. World energy still dependent heavily on Fossil Fuel (87%)

Ex. 4. Global Warming to cause massive economic and human losses

Health Care Cost Rise (1970 - 2007)



Energy Sources



Evidence of Supercomputing Populism

“The Recent Japanese NextGen SC Project Cancellation Fiasco”

- **May 2009 (precursor) – NEC&Hitachi announces pullout from Japanese NextGen SC Project**
 - **Result – public outcry denouncing NEC as “loser” “traitor”**
- **Sep. 2009 – new “Democratic Party” took over, slated to eliminate govt. waste**
- **Nov. 2009 – committees set-up, populated with “experts”(???) to review numerous projects and institutions out of the blue, just an hour each**
- **Nov. 16, 2009 – The NextGen project was reviewed, recommendation: “freeze (zero budget)” effectively killing the project(!)**
 - **“Why must we be #1? Why can’t we be just #2?”**



Retaliation of Scientists and the Public

- Immediate reaction by numerous academic societies – Physics, Informatics, Mechanical Engineering, ...
- Press conference by famous Nobel laureates denouncing government's decision
 - Head of Riken “Do the reviewers have the guts to be stand the trial of history”?
- Public outcry blaming the government for sacrificing science and engineering, endangering Japan's core competence
 - “It is worthless to aim for #2 in Science and Engineering”
- Due to public pressure, the project was resurrected
 - “SUPACON = Supercomputer” becomes a household terminology



Still we got our work cut out... Lack of Software Money and Talent



計算科学研究機構 (AICIS)。
それは、オールジャパン体制で臨む日本の計算科学の中心となる拠点。

10ペタフロップス級という世界最高水準の京速コンピュータ「京」を利用し、さまざまな分野における研究成果からブレイクスルーをもたらし、社会に貢献します。

また、次代を担う人材の育成にも積極的に取り組んでいきます。

**世界に誇る
拠点を**目指して

- トピックス** バックナンバーはこちら>>>
- 2010. 08. 12 神戸新聞ジュニア記者が来訪。
 - 2010. 07. 05 次世代スパコンの愛称が決定しました。
 - 2010. 07. 01 計算科学研究機構が発足しました。

ご挨拶>>>



計算科学研究機構 機構長
平尾 公彦

RIKEN WEB SITE
理化学研究所
ホームページ

次世代スパコン開発実施本部
ホームページ



独立行政法人 科学技術振興機構 募集TOP > 「ポストペタスケール高性能計算に資するシステムソフトウェア技術の創出」
「研究領域の概要」および「研究総括の募集・選考・研究領域運営にあたっての方針」

【CREST】

○戦略目標「メニーコアをはじめとした超並列計算環境に必要となるシステム制御等のための基盤的ソフトウェア技術の創出」の下の研究領域

ポストペタスケール高性能計算に資するシステムソフトウェア技術の創出

研究総括: 米澤 明憲(東京大学 大学院情報理工学系研究科 教授)

研究領域の概要

本研究領域は、次々世代(次世代スーパーコンピュータ「京」の次の世代)あるいはそれ以降のスーパーコンピュータに資する、システムソフトウェアやアプリケーション開発環境等の基盤技術の創出を目指すものです。

具体的には、2010年代半ば以降に多用される、メニーコア化された汎用型プロセッサや専用プロセッサ(現在GPGPUと呼ばれるものを含む)を用いて構成されるスーパーコンピュータの特徴を生かし、その上で実行されるアプリケーションを高効率・高信頼なものにするシステムソフトウェア(プログラミング言語、コンパイラ、ランタイムシステム、オペレーティングシステム、通信ミドルウェア、ファイルシステム等)、アプリケーション開発支援システム、超大規模データ処理システムソフトウェア等に関する、実用性を見据えた研究開発を対象とします。また、実用上の観点からそれらのソフトウェアレイアをまたがる研究開発が奨励されます。

[このページのトップへ](#)

研究総括の募集・選考・研究領域運営にあたっての方針

超大規模計算・記憶資源を活用した数値シミュレーションやデータ解析は、理論や実験・観測に加えて新たに登場した科学・技術の第三の方法論として、その役割の重要性が飛躍的に高まっています。これに呼応して、欧米、中国ではスーパーコンピュータの開発競争が激化し、我が国でも、2012年には次世代スーパーコンピュータ「京」の正式稼働が予定されています。スーパーコンピュータのこのような重要性に鑑み、各国でもすでに次の世代、すなわち次々世代のスーパーコンピュータの開発が水面下で進められ始めているのが現状です。

スーパーコンピュータの存在が有意義になるためには、その上で実行されるアプリケーション領域でのシミュレーションプログラムやデータ解析プログラムが開発されるのみならず、スーパーコンピュータのハードウェア性能を十分引き出す設計のもと高機能・高信頼性を有するシステムソフトウェアの存在が不可欠です。本研究領域プログラム、すなわち、プログラミング言語、コンパイラ、ランタイムシステム、オペレーティングシステム等や、アプリケーション開発支援システム(数値計算ライブラリを含む)、超大ソフトウェア等の研究開発を行います。

コンピュータのアーキテクチャーは、メニーコア化された汎用型プロセッサや専用プロセッサ(を含む)を用いて構成されるという方向性以外は、必ずしも明確になっているとは言えませ領域の研究課題の提案においては、研究開発で前提としているアーキテクチャーを出来る限り、採択された研究課題は、前提とするアーキテクチャー上で研究開発するシステムソフトウェア