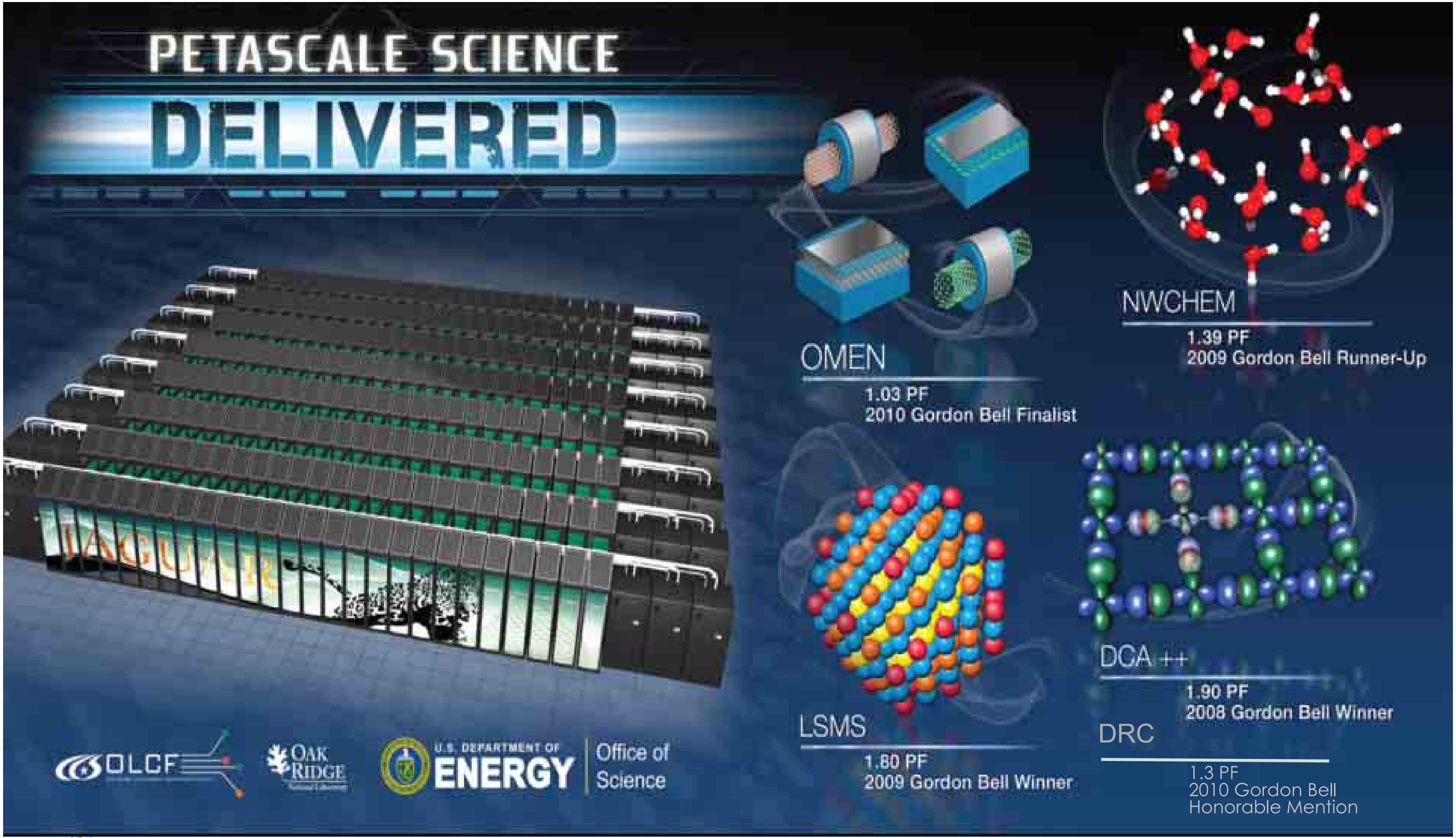
Petascale Applications



OMEN
1.03 PF
2010 Gordon Bell Finalist

NWCHEM
1.39 PF
2009 Gordon Bell Runner-Up

LSMS
1.80 PF
2009 Gordon Bell Winner

DCA ++
1.90 PF
2008 Gordon Bell Winner

DRC
1.3 PF
2010 Gordon Bell
Honorable Mention

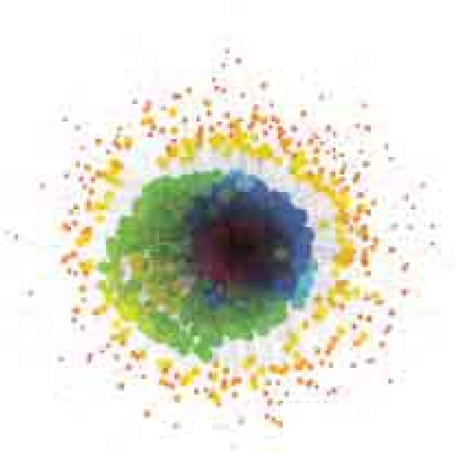
IBM in *Jeopardy*

- Watson – an AI system capable of answering questions posed in natural language
- IBM DeepQA software
- Cluster of Power750 systems
  - 2880 Power7 processors and 16 TB RAM
  - 3.5 GHz POWER 7 eight core processor, with four threads per core
  - Can process 500 Gigabytes per second
- Watson's data was stored on RAM rather than disk stores for faster access
  - 200 million pages of structured and unstructured data on 4 TB storage – including full text of Wikipedia.

Watson would use thousands of algorithms simultaneously to find answers

Graph 500: www.graph500.org

- First announced at ISC2010, first list released at SC2010
- The first serious approach to complement the Top 500 with data intensive applications
- Graph500 address 3 main graph kernels
  - concurrent search
  - optimization (single source shortest path)
  - edge-oriented (maximal independent set).
- The list addresses key application areas in Cybersecurity, Medical Informatics, Data Enrichment, Social Networks, and Symbolic Networks.
- Key Contributors: David A. Bader, Jonathan Berry, Simon Kahan, Richard Murphy, E. Jason Riedy, and Jeremiah Willcock.



Synthetic graph generated by a method called Kronecker multiplication

Sandia National Laboratories

Pacific Northwest NATIONAL LABORATORY

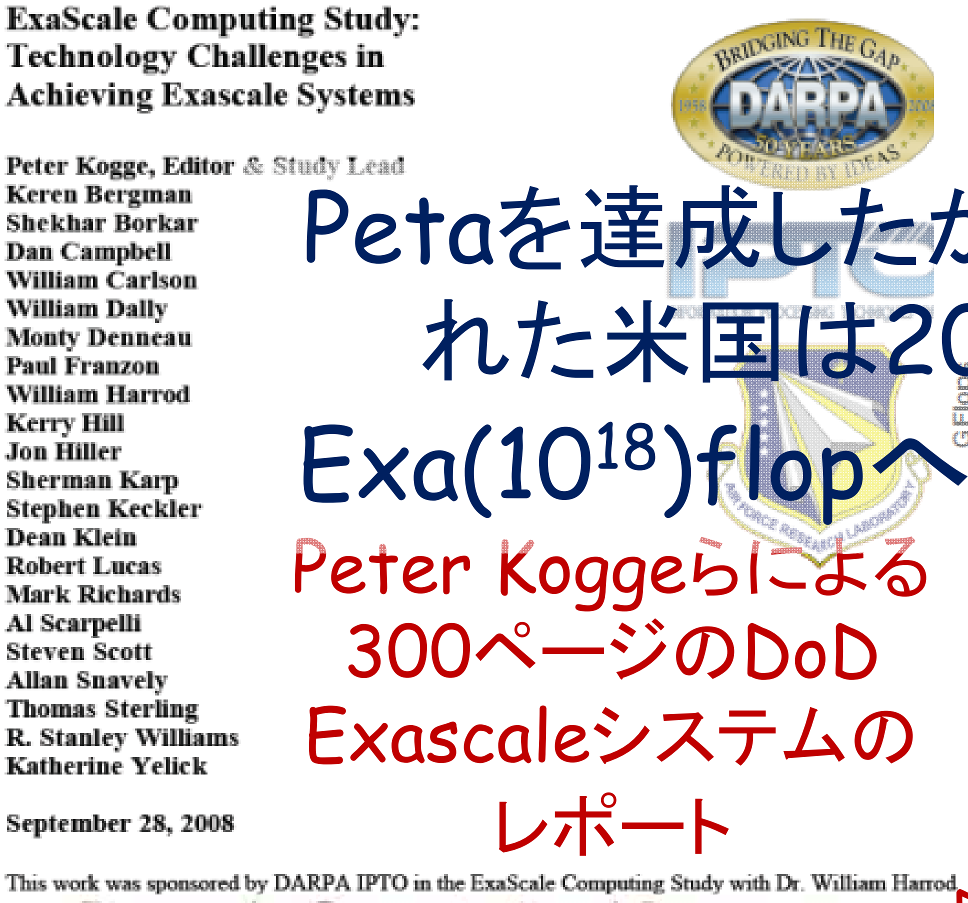
Georgia Tech

GRAPH 500

今後のペタ級マシン

Inst/Agency/Country(	Name	Machine	Peak Perf
ORNL/DoE/US	Jaguar Upgrade	Cray XT5	2.3PF
Tennessee大学/NSF/US 2009	Cracken	Cray XT5	1PF
Julich/欧州(ドイツ)	Jugene	IBM BG/P	1PF
中国・防衛大学	天河 (Tihanhe 1)	GPU Cluster/Dawning	1.2PF
中国・深圳国立スパコン	星雲 (Nebulae)	GPU Cluster/???	3PF
日本・東工大	TSUBAME2.0	GPU Cluster/HP-NEC	2.4 PF
LBNL/DoE/US 2010	Hopper	Cray XE6	1.3PF
中国・防衛大学	天河 (Tihanhe 1-A)	GPU Cluster/Dawning	5 PF
欧州PRACE計画・仏CEA	Tera 100	Nehalem-EX Cluster/Bull	1.25PF
ORNL/DoE/US	TITAN	Cray XK6 +GPU	20PF
NCSA/NSF/US	Blue Waters	IBM Power7 server	10PF
LLNL/DoE/US	Sequoia	IBM BG/Q	20PF
ArgonneNL/DoE/US 2011-12	Mira	IBM BG/Q	10PF
日本・理研	「京」	富士通 Venus 専用設計	10PF+
日本・筑波大	HA-PACS	GPU Cluster/HP-NEC	1PF
欧州ペタコン群/PRACE計画	???	IBM, Cray等	~PF x 4~5
中国	4~6個所	???Dawning?	合算数十PF以上

Petaを達成したが中国に抜かれた米国は2018-2020 Exa(10^{18})flopへ驀進を開始



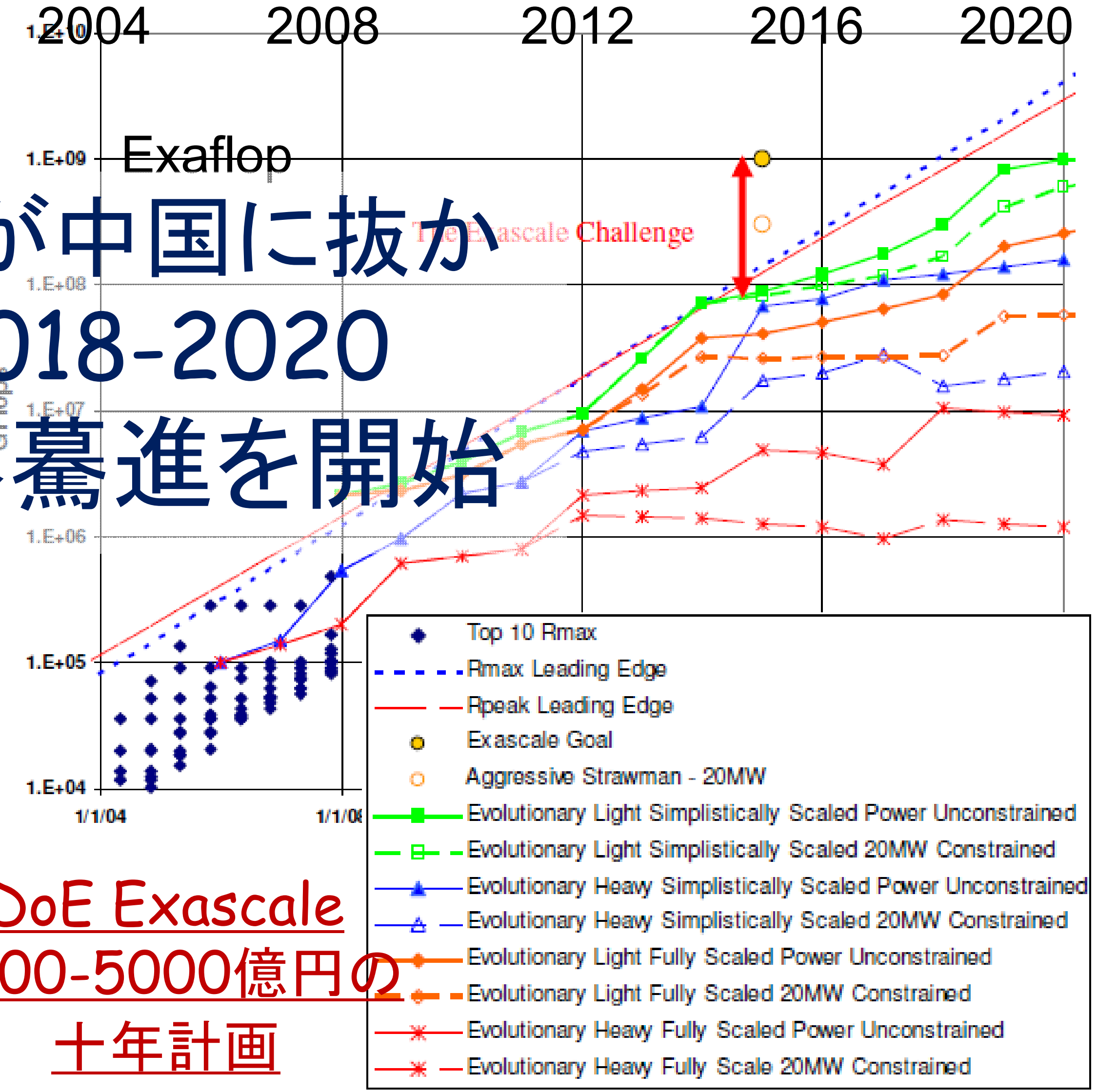
ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems

Peter Kogge, Editor & Study Lead
Keren Bergman
Shekhar Borkar
Dan Campbell
William Carlson
William Dally
Monty Denneau
Paul Franzon
William Harrod
Kerry Hill
Jon Hiller
Sherman Karp
Stephen Keckler
Dean Klein
Robert Lucas
Mark Richards
Al Scarpelli
Steven Scott
Allan Snavely
Thomas Sterling
R. Stanley Williams
Katherine Yelick

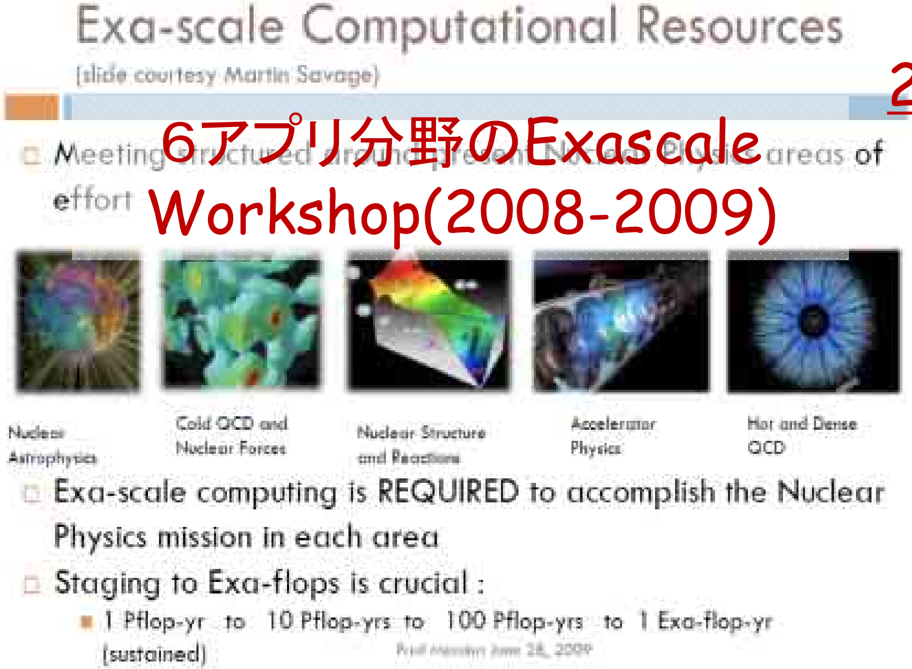
September 28, 2008

This work was sponsored by DARPA IPTO in the ExaScale Computing Study with Dr. William Harrod

Peter Koggeらによる300ページのDoD Exascaleシステムのレポート



DoE Exascale 2000-5000億円の十年計画



Exa-scale Computational Resources

[slide courtesy Martin Savage]

□ Meeting structured around present Nuclear Physics areas of effort

Nuclear Astrophysics
Cold QCD and Nuclear Forces
Nuclear Structure and Reactions
Accelerator Physics
Hot and Dense QCD

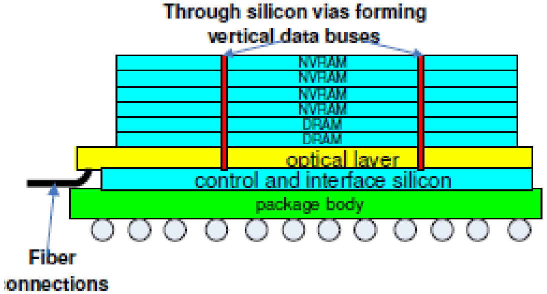
□ Exa-scale computing is REQUIRED to accomplish the Nuclear Physics mission in each area

□ Staging to Exa-flops is crucial :

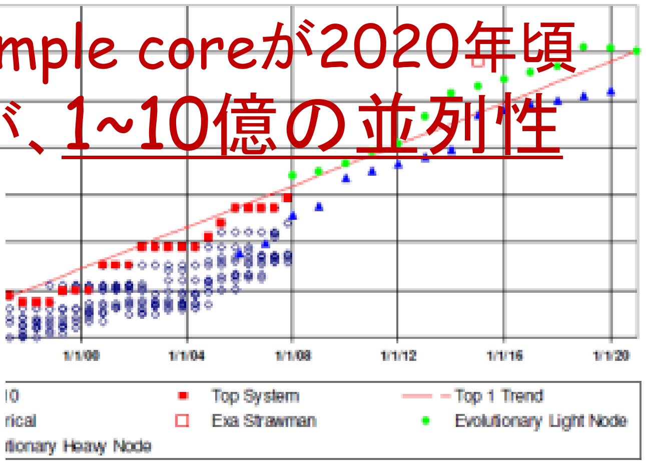
- 1 Pflop-yr to 10 Pflop-yrs to 100 Pflop-yrs to 1 Exa-flop-yr (sustained)

[illegible] June 28, 2009

6アプリ分野のExascale Workshop(2008-2009)



Through silicon vias forming vertical data buses
NVRAM
NVRAM
NVRAM
NVRAM
DRAM
DRAM
optical layer
control and interface silicon
package body
Fiber connections

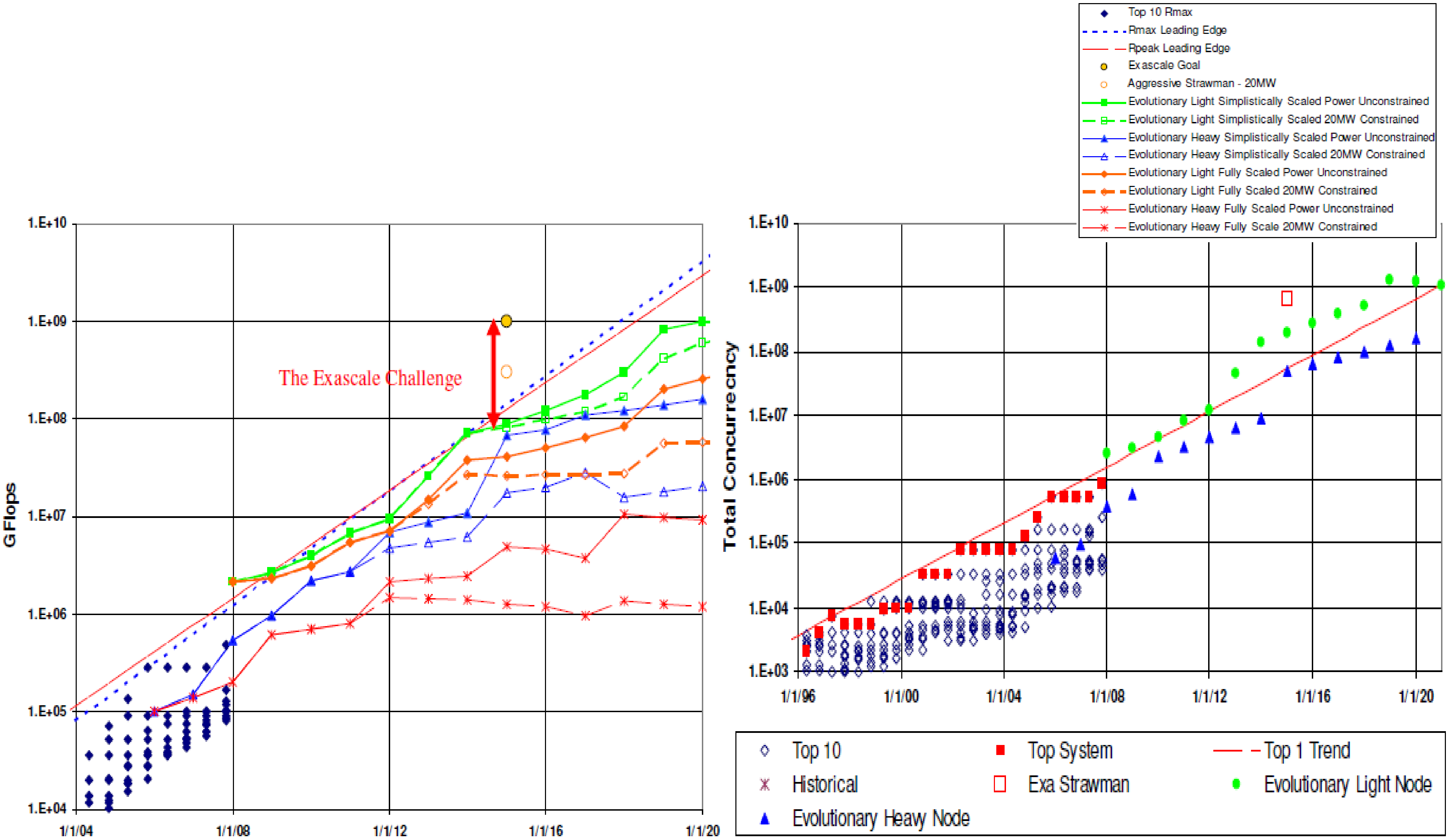


軽量なsimple coreが2020年頃有望だが、1~10億の並列性

ペタ～エクサへのスケーリングのロードブロック

- 「10億並列へ」は勇ましいが。。。
  - 電力･エネルギー
  - (強)スケーリングの欠落
  - N^2 vs. N 問題により深まるメモリ階層 (I/O 含む)
  - 極端に低まる信頼性と実行不能性
  - プログラミングや実行モデル

Extreme Many Core, Slow&Parallel is the Key to Low Power [Kogge08]



Case for Hybrid Multi/Many-Core Architectures

- Apps scaling governed by two components
  - Weak Scaling part, O(N) concurrency
  - Strong (Finite) Scaling Part, O(K) concurrency

N>>K for peta-exa

- S: speedup, H: fraction of O(N) execution, h: #weak-scale cores, f: #strong scale cores, c: speedup of strong scaling cores over weak scaling cores

$$S = \frac{1}{\frac{1-H}{fc} + \frac{H}{h}}$$

- 10PF machine analysis
  - Hybrid (TSUBAME2+): h=500,000, f=10,000, c=4, H=95% => S= 317,460 (Efficiency 63.4%)
  - Homo (BG/Q): h=800,000, f=10,000, c=1, H=95% => S= 161,616 (Efficiency 20.2%)
  - Gap will widen with Exascale, 1-10 billion cores

DoE: Reducing power is fundamentally about architecture choices & process technology

- Memory (2x-5x)
  - New memory interfaces (chip stacking and vias)
  - Replace DRAM with zero power non-volatile memory
- Processor (10x-20x)
  - New, power efficient architectures (many-core) and devices
  - Reducing data movement (functional reorganization, > 20x) with deep hierarchy, multi-chip layering & bonding
  - Domain/Core power gating and aggressive voltage scaling
- Interconnect (2x-5x)
  - More interconnect on package
  - Replace long haul copper with (redundunt) integrated optics
- Data Center Energy Efficiencies (10%-20%)
  - Higher operating temperature tolerance
  - Power supply and cooling efficiencies

DoE Exascale 性能指標

System attributes	“2010”		“2015”		“2018-20”	
System peak	2 PetaFlops		100-200 PetaFlops		1 ExaFlop	
Power	Jaguar 6 MW	TSUBAME 1.3 MW	15 MW		20 MW	
System Memory	0.3PB	0.1PB	5 PB		32-64PB	
Node Perf	125GF	1.6TF	0.5TF	7TF	1TF	10TF
Node Mem BW	25GB/s	0.5TB/s	0.1TB/s	1TB/s	0.4TB/s	4TB/s
Node Concurrency	12	O(1000)	O(100)	O(1000)	O(1000)	O(10000)
#Nodes	18,700	1442	50,000	5,000	1 million	100,000
Total Node Interconnect BW	1.5GB/s	8GB/s	20GB/s		200GB/s	
MTTI	O(days)		O(1 day)		O(1 day)	

Reliability and Resilience

- **Barriers**
  - Number of system components increasing faster than overall reliability
  - Silent error rates increasing
  - Reduced job progress due to fault recovery if we use existing checkpoint/restart
- **Technical Focus Areas**
  - Local recovery and migration
  - Development of a standard fault model and better understanding of types/rates of faults
  - Improved hardware and software reliability
    - Greater integration across entire stack
  - Fault resilient algorithms and applications
- **Technical Gap**
  - Maintaining today's MTTI given 10x - 100X increase in sockets will require:

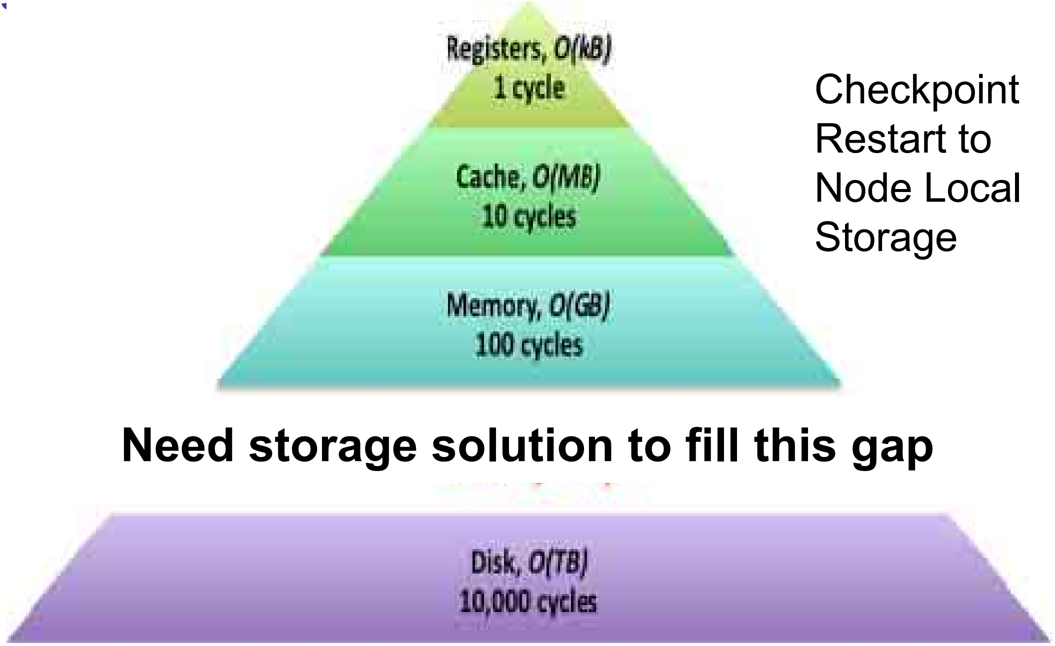
  10X improvement in hardware reliability

  10X in system software reliability, and

  10X improvement due to local recovery and migration as well as research in fault resilient applications
  - .

Taxonomy of errors (h/w or s/w)

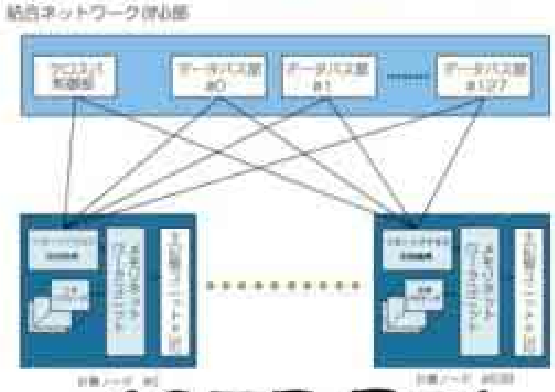
- **Hard errors**: permanent errors which cause system to hang or crash
- **Soft errors**: transient errors, either correctable or short term failure
- **Silent errors**: undetected errors either permanent or transient. *Concern is that simulation data or calculation have been corrupted and no error reported.*



Checkpoint Restart to Node Local Storage

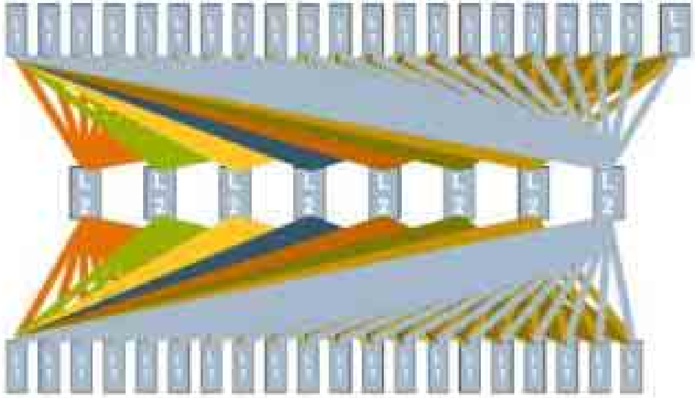
Need storage solution to fill this gap

Networks at 10 Petascale



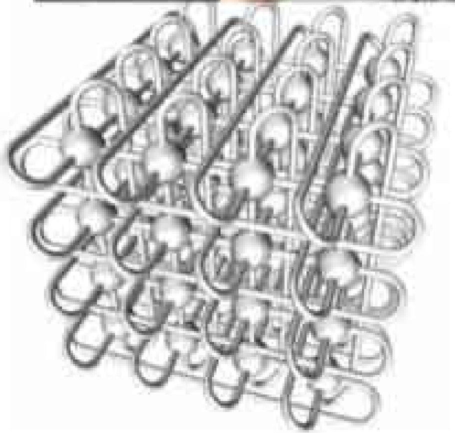
40TF ES1

12.8GB/s Link
5us latency
Full Crossbar
~12TB/s Bisection BW
3000km copper



10PF Ext. (5000 nodes) TSUBAME2.0

16GB/s node (QDR/EDRx4)
Less than 2us latency
Full Bisection Fat Tree
160TByte/s Bisection BW
1000km Fiber



10PF Kei

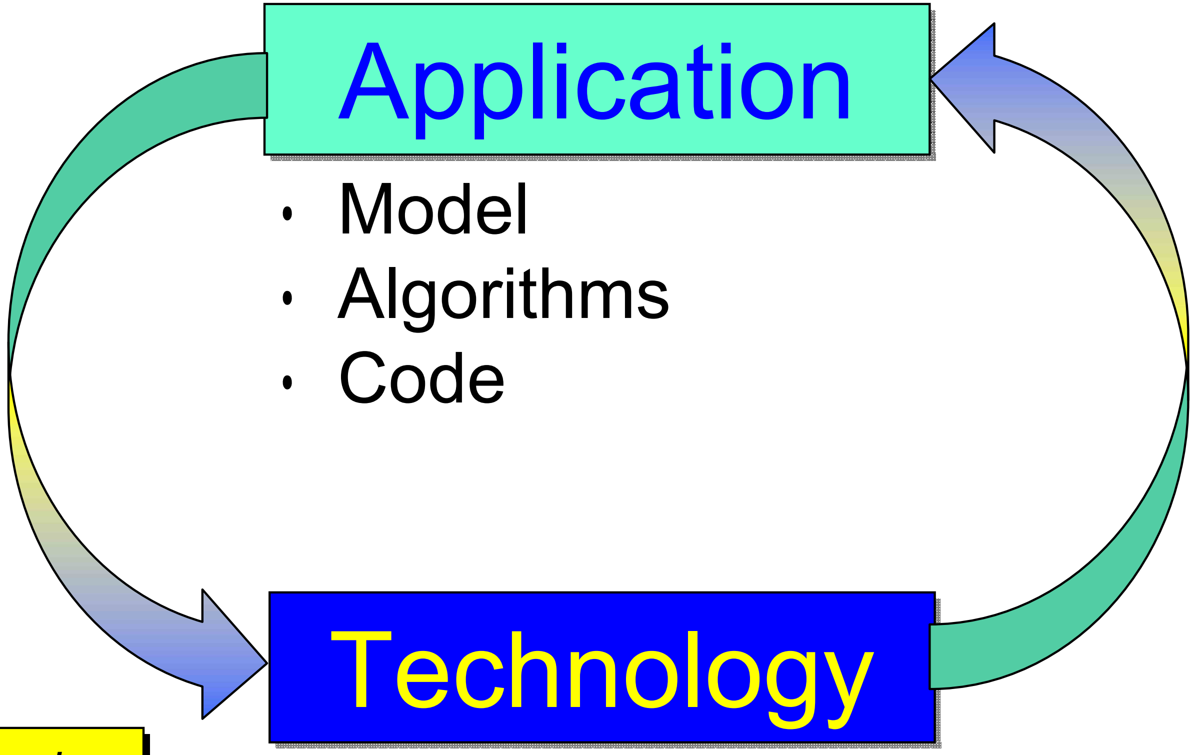
5GB/s Link
?us latency
6-D Torus
~30TByte/s?
Bisection BW

Interconnect Requirements

- Commodity solutions might not be sufficient=>HPC dedicated design (c.f. Infiniband)
- Very Low latency, Very High BW, very low power
  - High BW 200GB(Torus?)+ Low Latency (low radix) combo
  - stacking and/or other embedded design (c.f. BlueGene)
  - Global address space for low latency network
- Low cost & power=> redundant fiber-ribbon embedded VCSEL optics?
- Optical switching still difficult=> 200GB/s => At least 10Tbit/s switching capability for BW portion, < 1microsec latency for laltency portion

Co-design is a fundamental tenet of the initiative.

Application driven:
Find the best technology to run this code.
Sub-optimal



- Model
- Algorithms
- Code

⊕ architecture
⊕ programming model
⊕ resilience
⊕ power

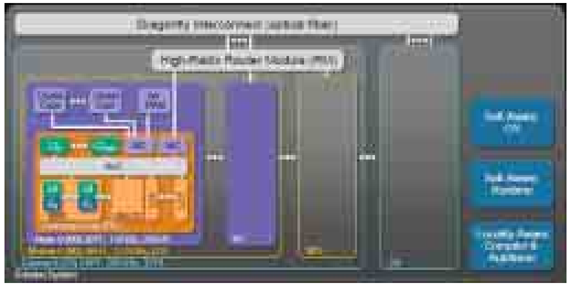
Technology driven:
Fit your application to this technology.
Sub-optimal.

Now, we must expand the co-design space to find better solutions:
•new applications & algorithms,
•better technology and performance.

DARPA UHPC

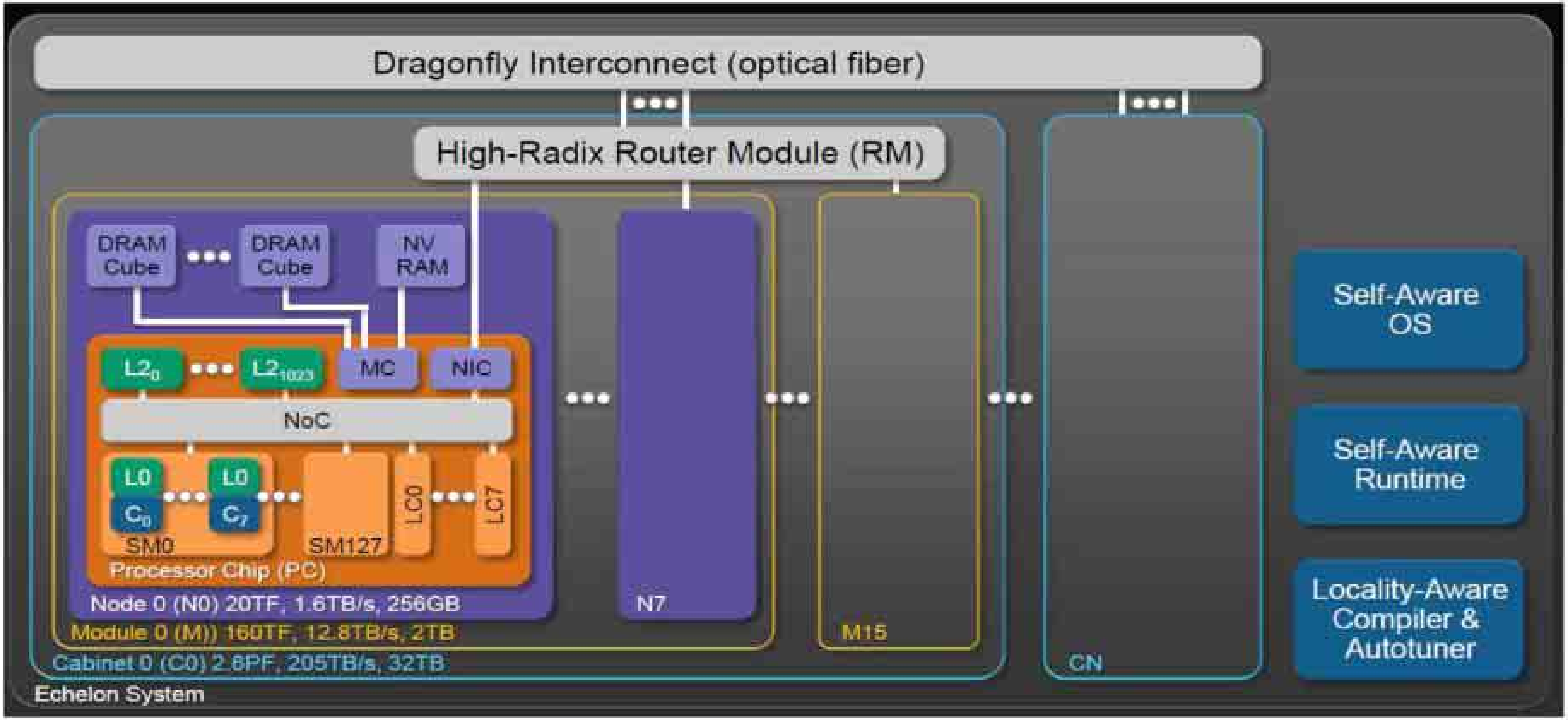
- DARPA Ubiquitous High Performance Computing (UHPC) program
  - Create an innovative, revolutionary new generation of exascale computing system
  - Design systems that overcome the limitations of current evolutionary approaches
- TA1: Systems Design and Implementation
  - 4 Awards : Sandia National Laboratory, Intel, NVIDIA, MIT CSAIL.
    - XCALIBER: SNL (Lead), Micron, LexisNexis, LSU, UIUC, UND, USC, UMD, GaTech, Stanford Univ,, NCSU
    - Runnymede – Intel, UDel., UIUC, UCSD, Reservoir Labs, ETI, SGI, Lockheed Martin, Cray
    - Echelon – NVIDIA, Cray, ORNL, and 6 top universities
    - MIT CSAIL
- TA2: Development of key metrics and benchmarks
  - Collaboration led by GaTech: participants include LLNL, LSU, MIT

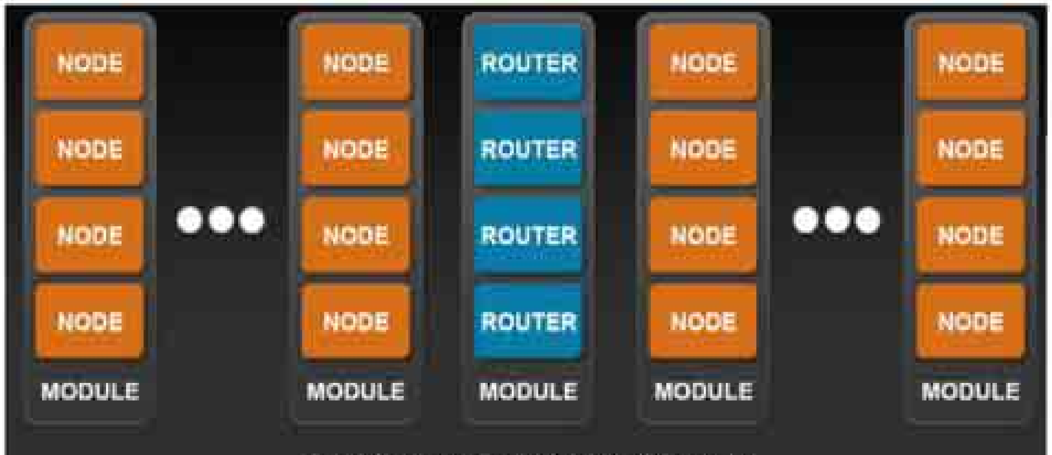
NVIDIA Echelon Architecture


Dragonfly Interconnect (optical fiber)
High-Radix Router Module (RM)
Self-Aware OS
Self-Aware Runtime
Locality-Aware Compiler & Autotuner

X-caliber Architecture

DARPA UHPC: NVIDIA Echelen

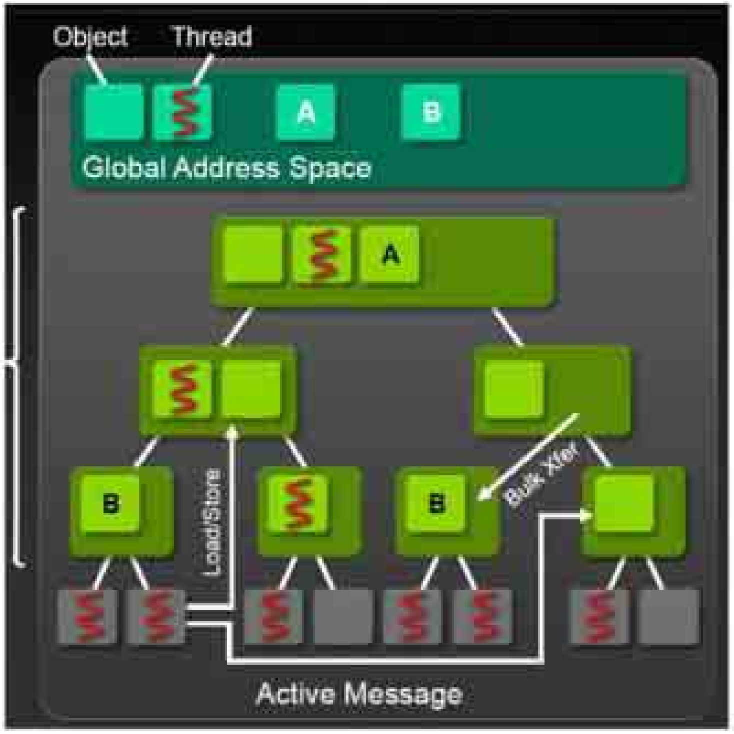

Dragonfly Interconnect (optical fiber)
High-Radix Router Module (RM)
DRAM Cube
DRAM Cube
NV RAM
$L2_0$
$L2_{1023}$
MC
NIC
NoC
L0
C_0
L0
C_7
SM0
SM127
LC0
LC7
Processor Chip (PC)
Node 0 (N0) 20TF, 1.6TB/s, 256GB
N7
Module 0 (M)) 160TF, 12.8TB/s, 2TB
M15
Cabinet 0 (C0) 2.6PF, 205TB/s, 32TB
CN
Echelon System
Self-Aware OS
Self-Aware Runtime
Locality-Aware Compiler & Autotuner

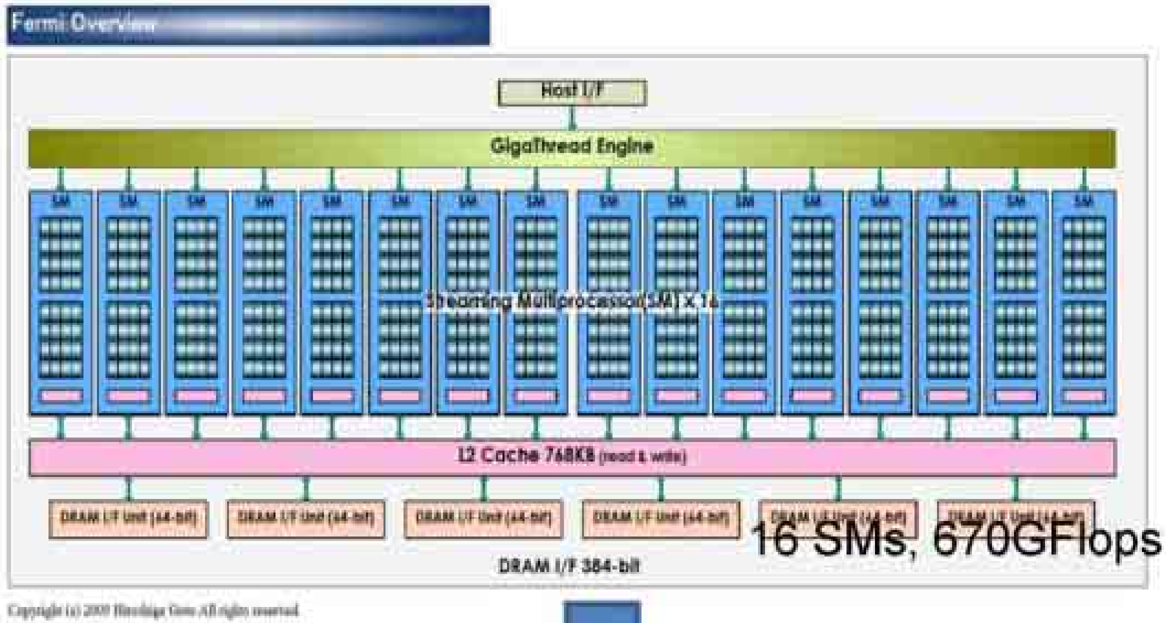


32 Modules, 4 Nodes/Module,
Central Router Module(s), Dragonfly Interconnect

Dragonfly Interconnect
400 Cabinets is ~1EF and ~15MW

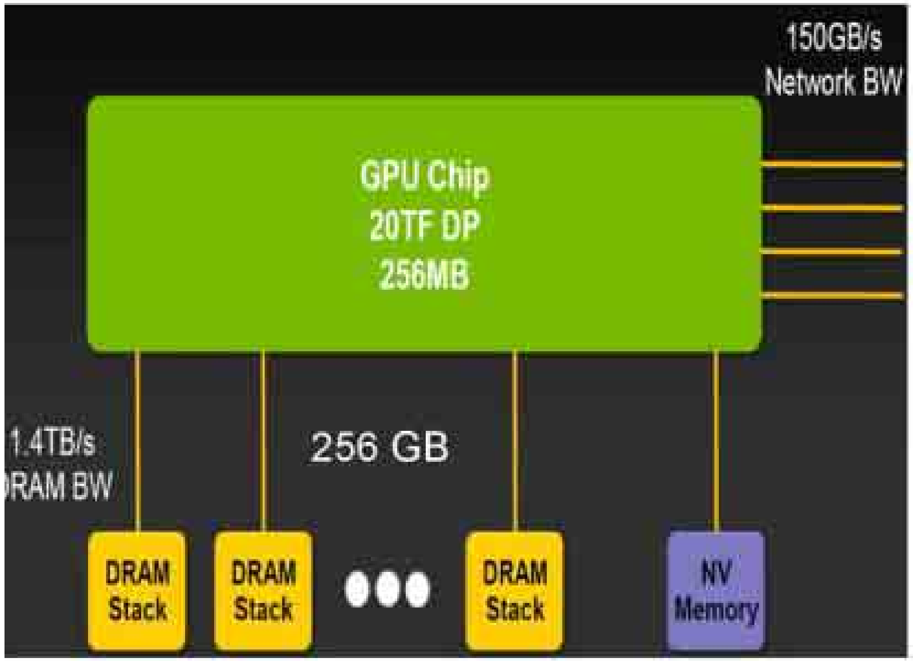
NVIDIA Echelon Processor&Network Design

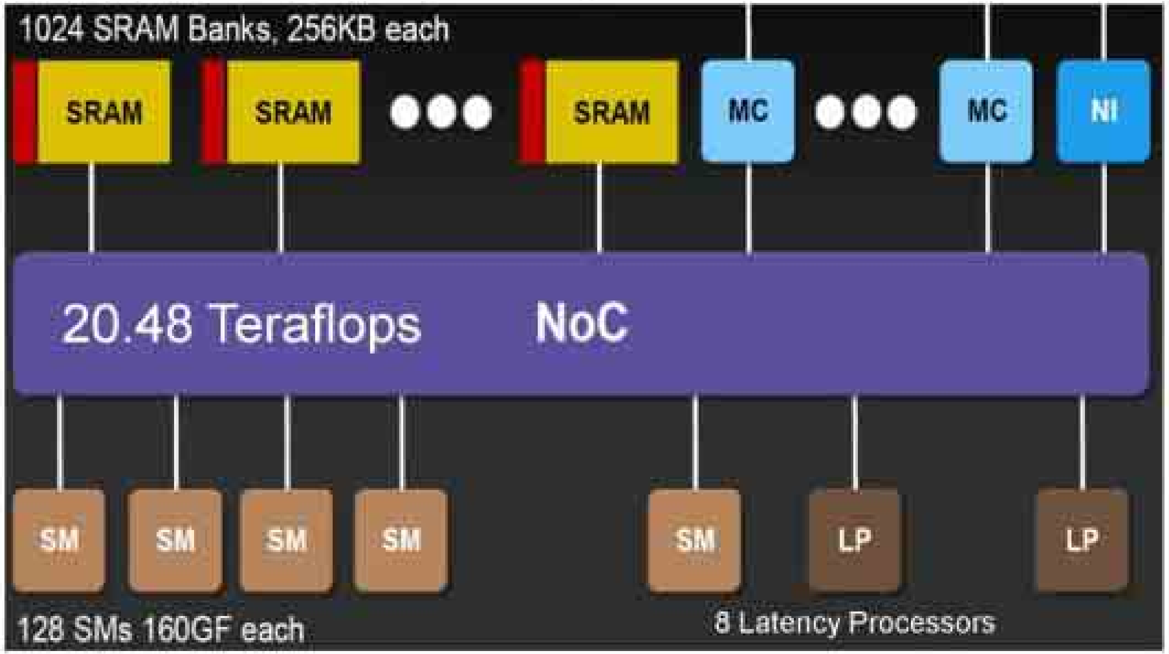




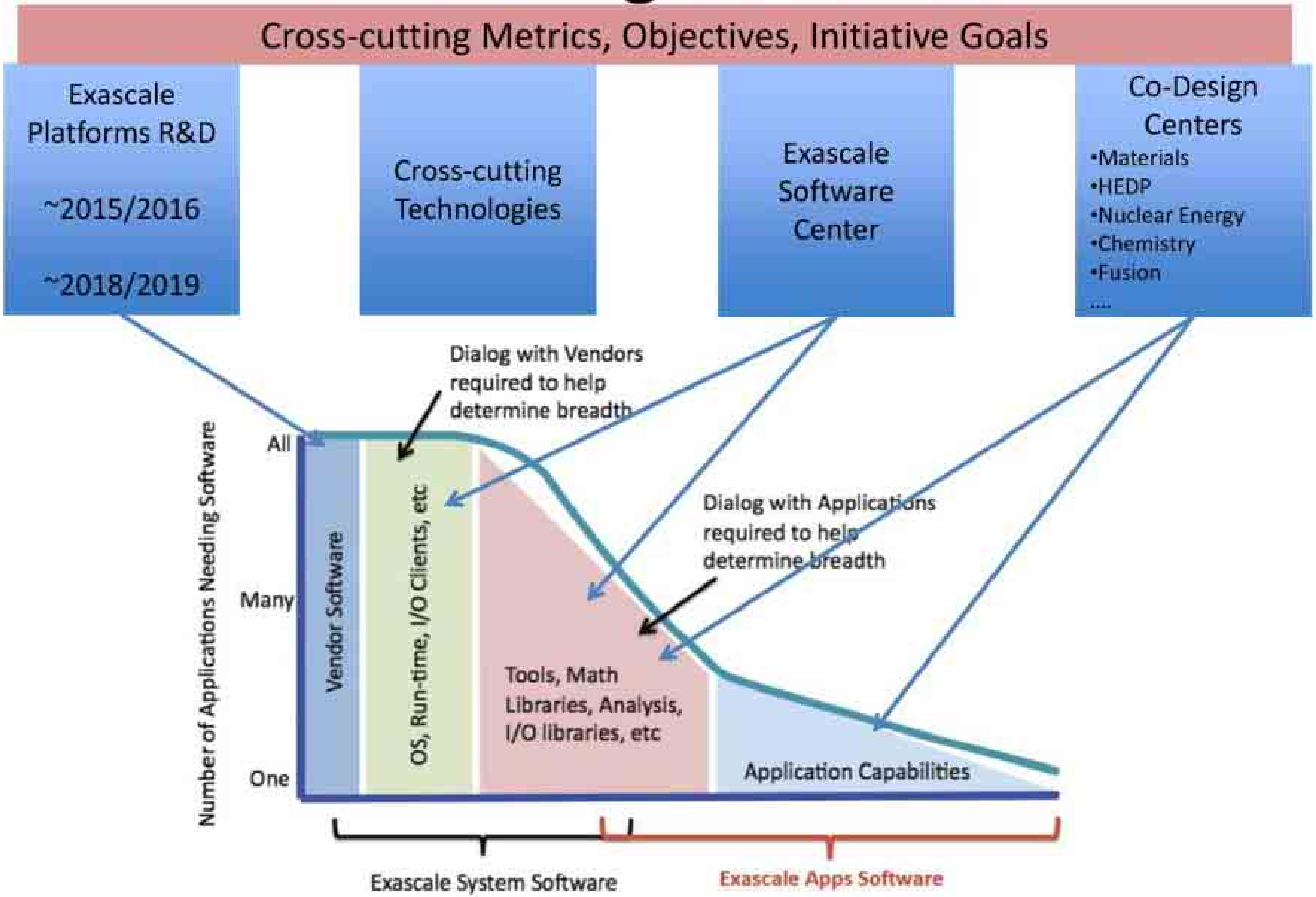
16 SMs, 670GFlops

Copyright (c) 2009 Hiroshige Goto All rights reserved





DoE Planning for Exascale



Cross-cutting Metrics, Objectives, Initiative Goals

Exascale Platforms R&D

~2015/2016

~2018/2019

Cross-cutting Technologies

Exascale Software Center

Co-Design Centers

- Materials
- HEDP
- Nuclear Energy
- Chemistry
- Fusion
- ....

Dialog with Vendors required to help determine breadth

Dialog with Applications required to help determine breadth

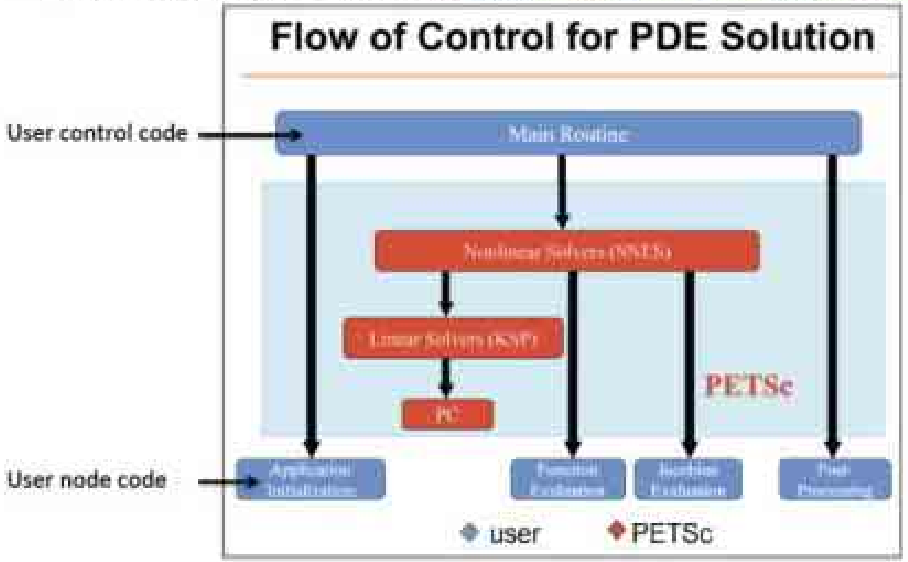
Number of Applications Needing Software
All
Many
One
Vendor Software
OS, Run-time, I/O Clients, etc
Tools, Math Libraries, Analysis, I/O libraries, etc
Application Capabilities
Exascale System Software
Exascale Apps Software

DoE Co-Design Centers

TITLE	LEAD PI
Center for Exascale Simulation of Advanced Reactors (CESAR)	Rosner (ANL)
FLASH High Energy Density Physics Exascale Codesign Center	Lamb (ANL)
The CERF Center: Co-design for Exascale Research in Fusion	Koniges (LBNL)
Exascale Co-Design Center for Materials in Extreme Environments: Engineering-Scale Predictions	Germann (LANL)
Chemistry Exascale Co-Design Center	Harrison (ORNL)
Combustion Exascale Co-Design Center	Chen (SNL)
Exascale Performance Research for Earth System Simulation (EXPRESS)	Jones (LANL)

Example of a SciDAC "petaskeleton"

- PETSc for PFLOTRAN http://ees.lanl.gov/pflotran/



- "Modernize" code for peta- to exaflops scalability
- "exaskeletons" as Co-Design Interface: sample implementations of basic"dynamical cores" (a step up from the 13"motifs")
- Skeletons should stress the HW and SW representatively; they should also be representative with respect to issues beyond FP performance, like checkpointing

The 13 algorithmic motifs*

- **Dense direct solvers**
- **Sparse direct solvers**
- **Spectral methods**
- **N-body methods**
- **Structured grids / iterative solvers**
- **Unstructured grids / iterative solvers**
- **Monte Carlo ("MapReduce")**
- **Combinatorial logic**
- **Graph traversal**
- **Graphical models**
- **Finite state machines**
- **Dynamic programming**
- **Backtrack and branch-and-bound**

* *The Landscape of Parallel Computing Research: The View from Berkeley, UCB/ EECS-2006-183*

EU Projects Towards Exascale

- **FP7-INFRA-2010.2.3.1 – First Implementation Phase of the European HPC service PRACE (PRACE1)**
  - Pan-European facilitation and operation of 1-10 Petaflops supercomputers (Tier-0)
- **FP7-INFRA-CSA(Coordination and Support Action) EESI (European Exascale Software Initiative)**
  - Create exascale software roadmap, contribute to IESP
- **FP7-INFRA-2010.2-RI- Structuring the European Research Area (PRACE2)**
  - Evolution of DEISA2, sub- to petaflop centers, direct coordination of governance and operations with PRACE1
- **FP7-ICT-2011.9.13 Exa-scale computing**
  - R&D Towards European Exascale SC, 2011-2014

FP7-ICT-2011.9.13 Exa-scale computing 25 million Euro, 2011-2014

- Must involve major PRACE centers and tech. vendors
- 60% systems (HW/SW), 40% apps, 2014 prototype deliverable
- 6 submitted, 3 accepted (注意：採択内容はまだ極秘)
- 1. Julich-Intel-others DEEP Project
  - Hybrid Intel MIC(Booster)-Cluster architecture, Extoll network, warm water cooling(45C), scalable hybrid software stack
- 2. BSC-ARM-others Mont-Blanc
  - Multi-Chip bonding of ~75 high-performance ARM Cortex A9 processors per socket + GPUs, next-gen 40-100Gbps Ethernet switch chips, hybrid execution model
- 3. HLRS-Cray CRESTA
  - All software: scalable next generation software stack and application scaling (e.g., ECMWF-xxx, GROMACS, …)---monitoring, autotuning, PGAS compilers,

- 日本の影は全くない