科学技術計算分科会 選出

SS 研 HPC フォーラム 2009 より

海外招待講演

Current Trends in High Performance Computing and Challenges for the Future

Jack Dongarra
University of Tennessee and
Oak Ridge National Laboratory

Current Trends in High Performance Computing and Challenges for the Future

Jack Dongarra
University of Tennessee and Oak Ridge National Laboratory

Abstract :

In this talk we examine how high performance computing has changed over the last 10-year and look toward the future in terms of trends. These changes have had and will continue to have a major impact on our numerical scientific software. A new generation of software libraries and algorithms are needed for the effective and reliable use of (wide area) dynamic, distributed and parallel environments. Some of the software and algorithm challenges have already been encountered, such as management of communication and memory hierarchies through a combination of compile -time and run- time techniques, but the increased scale of computation, depth of memory hierarchies, range of latencies, and increased run--time environment variability will make these problems much harder. We will focus on the redesign of software to fit multicore architectures.

Keyword :

High Performance Computing, Multicore, numerical software, algorithms, parallel computing

Profile:

Brief History :

Jack Dongarra received a Bachelor of Science in Mathematics from Chicago State University in 1972 and a Master of Science in Computer Science from the Illinois Institute of Technology in 1973. He received his Ph.D. in Applied Mathematics from the University of New Mexico in 1980. He worked at the Argonne National Laboratory until 1989, becoming a senior scientist. He now holds an appointment as University Distinguished Professor of Computer Science in the Computer Science Department at the University of Tennessee and holds the title of Distinguished Research Staff in the Computer Science and Mathematics Division at Oak Ridge National Laboratory (ORNL), Turing Fellow at Manchester University, and an Adjunct Professor in the Computer Science Department at Rice University. He is the director of the Innovative Computing Laboratory at the University of Tennessee. He is also the director of the Center for Information Technology Research at the University of Tennessee which coordinates and facilitates IT research efforts at the University.

Field of research :

He specializes in numerical algorithms in linear algebra, parallel computing, the use of advanced-computer architectures, programming methodology, and tools for parallel computers. His research includes the development, testing and documentation of high quality mathematical software. He has contributed to the design and implementation of the following open source software packages and systems: EISPACK, LINPACK, the BLAS, LAPACK, ScaLAPACK, Netlib, PVM, MPI, NetSolve, Top500, ATLAS, and PAPI. He has published approximately 200 articles, papers, reports and technical memoranda and he is coauthor of several books.

Academic society, Award, Book, etc. :

H He was awarded the IEEE Sid Fernbach Award in 2004 for his contributions in the application of high performance computers using innovative approaches and in 2008 he was the recipient of the first IEEE Medal of Excellence in Scalable Computing. He is a Fellow of the AAAS, ACM, IEEE, and SIAM and a member of the National Academy of Engineering.

An Overview of High Performance Computing and Future Requirements

Jack Dongarra

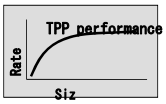
University of Tennessee
Oak Ridge National Laboratory

TOP500

H. Meuer, H. Simon, E. Strohmaier, & JD

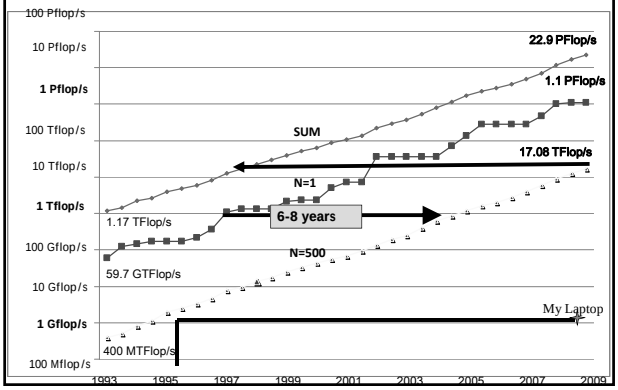
- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP

  $Ax=b$, *dense problem*

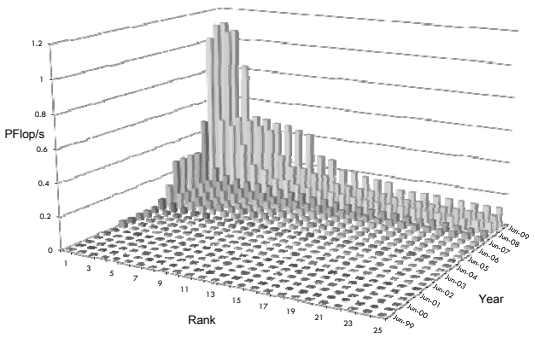


- Updated twice a year
  SC'xy in the States in November
  Meeting in Germany in June
- All data available from **www.top500.org**

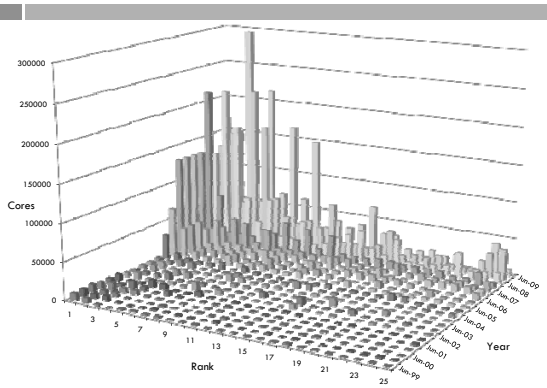
Performance Development



Performance of Top25 Over 10 Years



Cores in the Top25 Over Last 10 Years



Looking at the Gordon Bell Prize

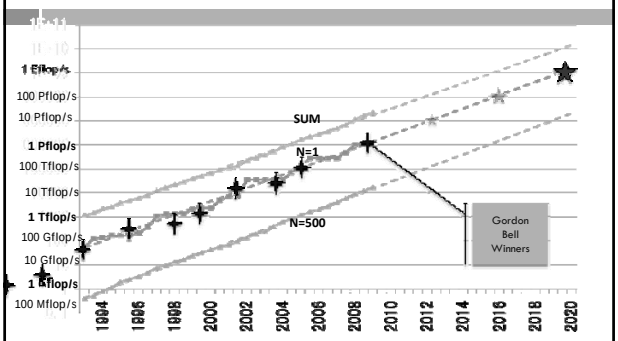
(Recognize outstanding achievement in high-performance computing applications and encourage development of parallel processing)

- 1 GFlop/s; 1988; Cray Y-MP; 8 Processors
  - Static finite element analysis
- 1 TFlop/s; 1998; Cray T3E; 1024 Processors
  - Modeling of metallic magnet atoms, using a variation of the locally self-consistent multiple scattering method.

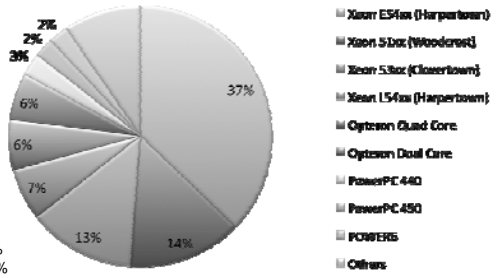
- 1 PFlop/s; 2008; Cray XT5; 1.5×10^5 Processors
  - Superconductive materials

- 1 EFlop/s; ~2018; ?; 1×10^7 Processors (10^9 threads)

Performance Development in Top500

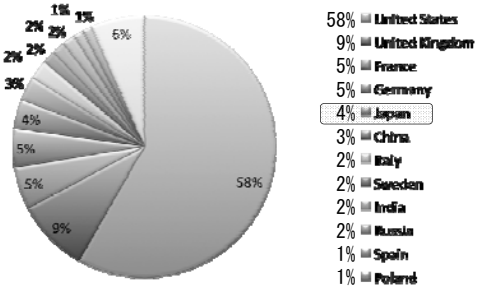


Processors Used in Supercomputers



Intel 71%
AMD 13%
IBM 7%

Countries / System Share

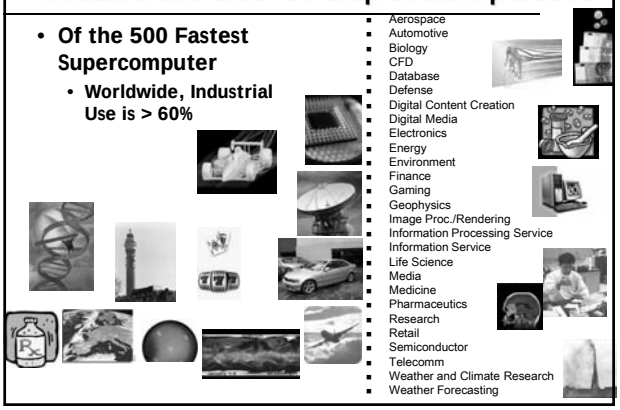


58% United States
9% United Kingdom
5% France
5% Germany
4% Japan
3% China
2% Italy
2% Sweden
2% India
2% Russia
1% Spain
1% Poland

Industrial Use of Supercomputers

- **Of the 500 Fastest Supercomputer**
  - **Worldwide, Industrial Use is > 60%**

- Aerospace
- Automotive
- Biology
- CFD
- Database
- Defense
- Digital Content Creation
- Digital Media
- Electronics
- Energy
- Environment
- Finance
- Gaming
- Geophysics
- Image Proc./Rendering
- Information Processing Service
- Information Service
- Life Science
- Media
- Medicine
- Pharmaceutics
- Research
- Retail
- Semiconductor
- Telecomm
- Weather and Climate Research
- Weather Forecasting



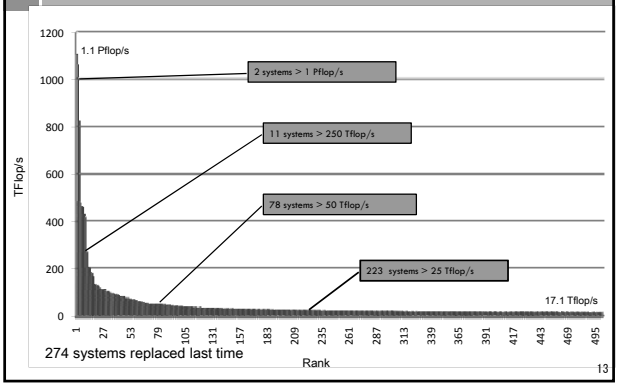
33rd List: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Tflops]	% of Peak
1	DOE / NNSA Los Alamos Nat Lab	Roadrunner / IBM BladeCenter QS22/LS21	USA	129,600	1,105	76
2	DOE / OS Oak Ridge Nat Lab	Jaguar / Cray Cray XT5 QC 2.3 GHz	USA	150,152	1,059	77
3	Forschungszentrum Juelich (FZJ)	Jugene / IBM Blue Gene/P Solution	Germany	294,912	825	82
4	NASA / Ames Research Center/NAS	Pleiades / SGI SGI Altix ICE 8200EX	USA	51,200	480	79
5	DOE / NNSA Lawrence Livermore NL	BlueGene/L IBM eServer Blue Gene Solution	USA	212,992	478	80
6	NSF NICS/U of Tennessee	Kraken / Cray Cray XT5 QC 2.3 GHz	USA	66,000	463	76
7	DOE / OS Argonne Nat Lab	Intrepid / IBM Blue Gene/P Solution	USA	163,840	458	82
8	NSF TACC/U. of Texas	Ranger / Sun SunBlade x6420	USA	62,976	433	75
9	DOE / NNSA Lawrence Livermore NL	Dawn / IBM Blue Gene/P Solution	USA	147,456	415	83
10	Forschungszentrum Juelich (FZJ)	JUROPA /Sun - Bull SA NovaScale /Sun Blade	Germany	26,304	274	89

33rd List: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Tflops]	% of Peak	Power [MW]	Flops/ Watt
1	DOE / NNSA Los Alamos Nat Lab	Roadrunner / IBM BladeCenter QS22/LS21	USA	129,600	1,105	76	2.48	446
2	DOE / OS Oak Ridge Nat Lab	Jaguar / Cray Cray XT5 QC 2.3 GHz	USA	150,152	1,059	77	6.95	151
3	Forschungszentrum Juelich (FZJ)	Jugene / IBM Blue Gene/P Solution	Germany	294,912	825	82	2.26	365
4	NASA / Ames Research Center/NAS	Pleiades / SGI SGI Altix ICE 8200EX	USA	51,200	480	79	2.09	230
5	DOE / NNSA Lawrence Livermore NL	BlueGene/L IBM eServer Blue Gene Solution	USA	212,992	478	80	2.32	206
6	NSF NICS/U of Tennessee	Kraken / Cray Cray XT5 QC 2.3 GHz	USA	66,000	463	76		
7	DOE / OS Argonne Nat Lab	Intrepid / IBM Blue Gene/P Solution	USA	163,840	458	82	1.26	363
8	NSF TACC/U. of Texas	Ranger / Sun SunBlade x6420	USA	62,976	433	75	2.0	217
9	DOE / NNSA Lawrence Livermore NL	Dawn / IBM Blue Gene/P Solution	USA	147,456	415	83	1.13	367
10	Forschungszentrum Juelich (FZJ)	JUROPA /Sun - Bull SA NovaScale /Sun Blade	Germany	26,304	274	89	1.54	178

Distribution of the Top500



15 Systems on Top 500 in Japan

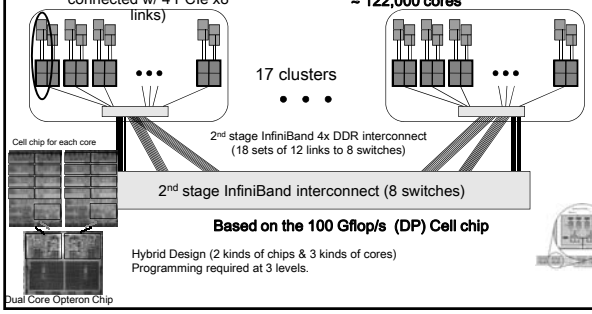
Rank	Location	Manuf	Machine	Year	Cores	Rmax	Rpeak
22	The Earth Simulator Center	NEC	Earth Simulator	2009	1280	122400	131072
28	JAXA	Fujitsu	Fujitsu FX1, Quadcore SPARC64 VII 2.52 GHz, Infiniband DDR	2009	12032	110600	121282
40	Inst of Phy & Chem. Res (RIKEN)	Fujitsu	PRIMERGY RX200S5 Cluster, Xeon X5570 2.93GHz, Infiniband DDR	2009	8256	87890	96760
41	GSIC Center, TiTech	NEC/Sun	Sun Fire x4600/x6250, AMD 2.4/2.6 GHz, Xeon E5440 2.833 GHz, ClearSpeed CSX600, nVidia GT200; Voltaire	2009	31024	87010	163188
42	Information Tech Center, U of Tokyo	Hitachi	Hitachi Cluster Opteron QC 2.3 GHz, Myrinet 10G	2008	12288	82984	113050
47	CCS, U of Tsukuba	Appro	Appro Xtreme-X3 Server - Quad Opteron Quad Core 2.3 GHz, Infiniband	2009	10368	77280	95385
65	Nat Inst for Fusion Science (NIFS)	Hitachi	Hitachi SR16000 Model L2, Power6 4.7Ghz, Infiniband	2009	4096	56650	77004
69	U of Tokyo/Human Genome C, IMS	Sun	SunBlade x6250, Xeon E5450 3GHz, Infiniband	2009	5760	54210	69120
78	Kyoto University	Fujitsu	Fujitsu Cluster HX600, Opteron Quad Core, 2.3 GHz, Infiniband	2008	6656	50510	61235
93	National Inst for Materials Science	SGI	SGI Altix ICE 8200EX, Xeon X5560 quad core 2.8 GHz	2009	4096	42690	45875
259	Nat Astro Obs of Japan	Cray Inc.	Cray XT4 QuadCore 2.2 GHz	2008	3248	22930	28582
277	Nat Astro Obs of Japan	Self-made	GRAPE-DR accelerator Cluster, Infiniband	2009	8192	21960	84480
394	Comp Biology Res Center, AIST	IBM	eServer Blue Gene Solution	2005	8192	18665	22937
397	High Energy Acc Research Org /KEK	IBM	eServer Blue Gene Solution	2006	8192	18665	22937
398	High Energy Acc Research Org /KEK	IBM	eServer Blue Gene Solution	2006	8192	18665	22937

LANL Roadrunner
A Petascale System in 2008

"Connected Unit" cluster
192 Opteron nodes
(180 w/ 2 dual-Cell blades
connected w/ 4 PCIe x8 links)

≈ 13,000 Cell HPC chips
≈ **1.33 PetaFlop/s** (from Cell)
≈ 7,000 dual-core Opterons
≈ **122,000 cores**

17 clusters



Based on the 100 Gflop/s (DP) Cell chip

Hybrid Design (2 kinds of chips & 3 kinds of cores)
Programming required at 3 levels.

ORNL's Newest System Jaguar XT5

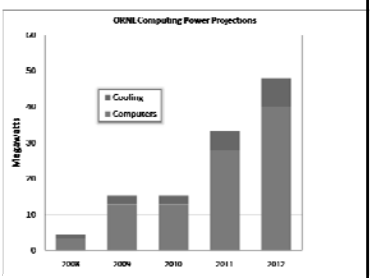
Jaguar	Total	XT5	XT4
Peak Performance	**1,645**	**1,382**	**263**
AMD Opteron Cores	181,504	150,176	31,328
System Memory (TB)	362	300	62
Disk Bandwidth (GB/s)	284	240	44
Disk Space (TB)	10,750	10,000	750
Interconnect Bandwidth (TB/s)	532	374	157

Will be upgraded this year to a 2 Pflop/s system with > 224K AMD Istanbul Cores.

U.S. DEPARTMENT OF ENERGY | Office of Science

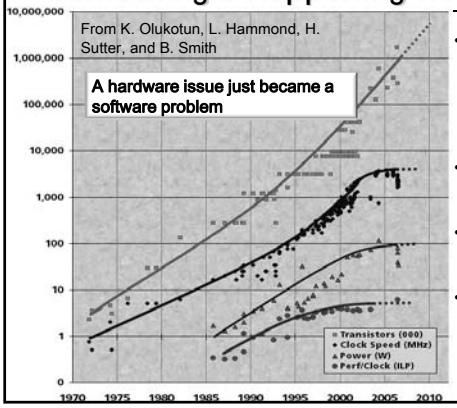
ORNL/UTK Computer Power Cost Projections 2008-2012

- Over the next 5 years ORNL/UTK will deploy 2 large Petascale systems
- Using 15 MW today
- By 2012 close to 50MW!!
- Power costs close to $10M today.
- Cost estimates based on $0.07 per KwH



Power becomes the architectural driver for future large systems

Something's Happening Here…



- **In the "old days" it was: each year processors would become faster**
- **Today the clock speed is fixed or getting slower**
- **Things are still doubling every 18 -24 months**
- **Moore's Law reinterpretated.**
  - **Number of cores double every 18-24 months**

Moore's Law Reinterpreted

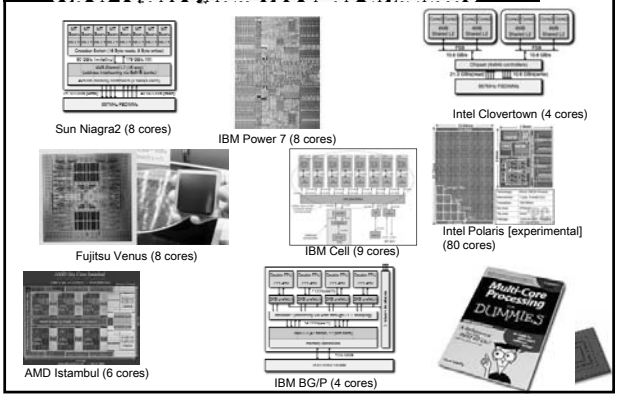
- **Number of cores per chip doubles every 2 year, while clock speed decreases (not increases).**
  - **Need to deal with systems with millions of concurrent threads**
    - **Future generation will have billions of threads!**
  - **Need to be able to easily replace inter-chip parallelism with intro-chip parallelism**
- **Number of threads of execution doubles every 2 year**

19

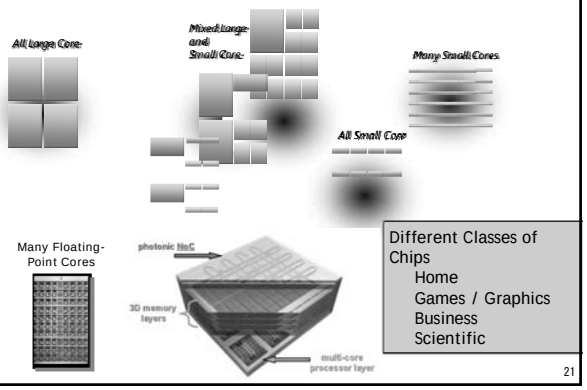
Today's Multicores

99% of Top500 Systems Are Based on Multicore

282 use Quad-Core
204 use Dual-Core
3 use Nona-core



What's Next?



Different Classes of Chips
- Home
- Games / Graphics
- Business
- Scientific

21

Commodity

- **Moore's "Law" favored consumer commodities**
  - **Economics drove enormous improvements**
  - **Specialized processors and mainframes faltered**
  - **Custom HPC hardware largely disappeared**
  - **Hard to compete against 50%/year improvement**
- **Implications**
  - **Consumer product space defines outcomes**
  - **It does not always go where we hope or expect**
  - **Research environments track commercial trends**
  - **Driven by market economics**
  - **Think about processors, clusters, commodity storage**

22

Future Computer Systems

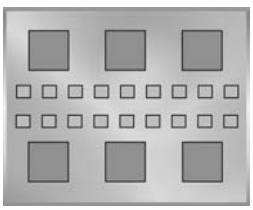
- **Most likely be a hybrid design**
- **Think standard multicore chips and accelerator (GPUs)**
- **Today accelerators are attached**
- **Next generation more integrated**
- **Intel's Larrabee in 2010**
  - **8,16,32,or 64 x86 cores**
- **AMD's Fusion in 2011**
  - **Multicore with embedded graphics ATI**
- **Nvidia's plans?**

Intel Larrabee

23

Architecture of Interest

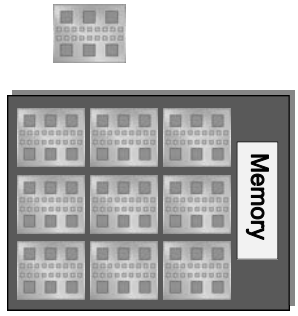
- **Manycore chip**
- **Composed of hybrid cores**
  - **Some general purpose**
  - **Some graphics**
  - **Some floating point**



24

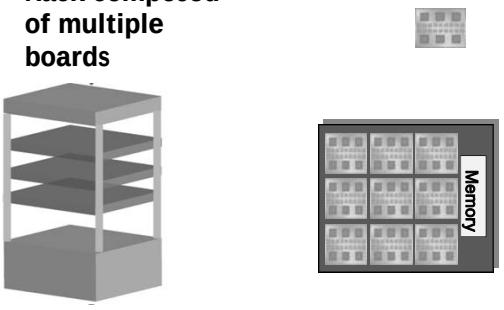
Architecture of Interest

- Board composed of multiple chips sharing memory



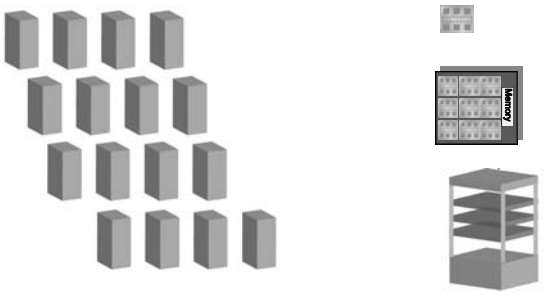
Architecture of Interest

- Rack composed of multiple boards



Architecture of Interest

- A room full of these racks



- Think millions of cores

Major Changes to Software

- Must rethink the design of our software
  - Another disruptive technology
    - Similar to what happened with cluster computing and message passing
  - Rethink and rewrite the applications, algorithms, and software
- Numerical libraries for example will change
  - For example, both LAPACK and ScaLAPACK will undergo major changes to accommodate this

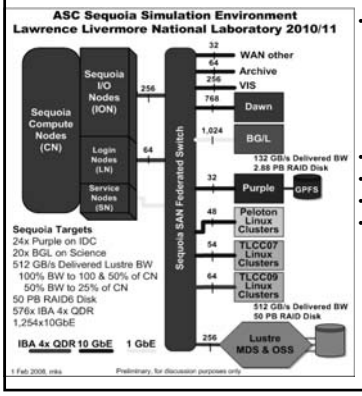
Quasi Mainstream Programming Models

- C, Fortran, C++ and MPI
- OpenMP, pthreads
- (CUDA, RapidMind, Cn) → OpenCL
- PGAS (UPC, CAF, Titanium)
- HPCS Languages (Chapel, Fortress, X10)
- HPC Research Languages and Runtime
- HLL (Parallel Matlab, Grid Mathematica, etc.)

DOE Office of Science Next System

- DOE's requirement for 20-40 PF of compute capability split between the Oak Ridge and Argonne LCF centers
- ORNL's proposed system will be based on accelerator technology includes software development environment
- Plans are to deploy the system in late 2011 with users getting access in 2012

Sequoia LLNL



- Diverse usage models drive platform and simulation environment requirements
  - Will be 2D ultra-res and 3D high-res Quantification of Uncertainty engine
  - 3D Science capability for known unknowns and unknown unknowns
- Peak 20 petaFLOP/s
- IBM BG/Q
- Target production 2011-2016
- Sequoia Component Scaling
  - Memory B:F = 0.08
  - Mem BW B:F = 0.2
  - Link BW B:F = 0.1
  - Min Bisect B:F = 0.03
  - SAN BW GB/:PF/s = 25.6
  - F is peak FLOP/s

31

NSF University of Illinois; Blue Waters

Blue Waters will be the powerhouse of the National Science Foundation's strategy to support supercomputers for scientists nationwide

	Blue Waters	NCSA/Illinois	1 Pflop *sustained* per second
T1			
T2	Ranger	TACC/U of Texas	504 Tflop/s peak per second
T2	Kraken	NICS/U of Tennessee	1 Pflops peak per second
T2	Campuses across the U.S.	Several sites	50-100 Tflops peak per second
T3			

32

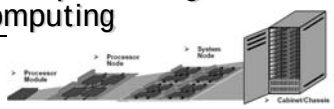
NSF University of Illinois; Blue Waters

Blue Waters - Main Characteristics

- **Hardware:**
  - **Processor: IBM Power7 multicore architecture**
    - **8 cores per chip**
    - **32 Gflop/s per core; 256 Gflop/s chip**
  - **More than 200,000 cores will be available**
  - **Capable of simultaneous multithreading (SMT)**
  - **Vector multimedia extension capability (VMX)**
  - **Four or more floating-point operations per cycle**
  - **Multiple levels of cache – L1, L2, shared L3**
  - **32 GB+ memory per SMP, 2 GB+ per core**
  - **16+ cores per SMP**
  - **10+ Petabytes of disk storage**
  - **Network interconnect with RDMA technology**

33

DARPA's RFI on Ubiquitous High Performance Computing



- *1 PFLOP/S HPL, air-cooled, single 19-inch cabinet ExtremeScale system.*
- *The power budget for the cabinet is 57 kW, including cooling.*
- *Achieve 50 GFLOPS/W for the High-Performance Linpack (HPL) benchmark.*
- *The system design should provide high performance for scientific and engineering applications.*
- *The system should be a highly programmable system that does not require the application developer to directly manage the complexity of the system to achieve high performance.*
- *The system must explicitly show a high degree of innovation and software and hardware co-design throughout the life of the program.*
- *5 phases;*
  - *1) conceptual designs; 2) execution model; 3) full-scale simulation; 4) delivery; 5) modify and refine.*

34

Exascale Computing

- **Exascale systems are likely feasible by 2017±2**
- **10-100 Million processing elements (cores or mini-cores) with chips perhaps as dense as 1,000 cores per socket, clock rates will grow more slowly**
- **3D packaging likely**
- **Large-scale optics based interconnects**
- **10-100 PB of aggregate memory**
- **Hardware and software based fault management**
- **Heterogeneous cores**
- **Performance per watt — stretch goal 100 GF/watt of sustained performance ⇒ >> 10 - 100 MW Exascale system**
- **Power, area and capital costs will be significantly higher than for today's fastest systems**

Google: exascale computing study

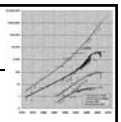
35

IESP: The Need

- **The largest scale systems are becoming more complex, with designs supported by consortium**
  - **The software community has responded slowly**
- **Significant architectural changes evolving**
  - **Software must dramatically change**
- **Our ad hoc community coordinates poorly, both with other software components and with the vendors**
  - **Computational science could achieve more with improved development and coordination**

36

A Call to Action



- Hardware has changed dramatically while software ecosystem has remained stagnant
- Previous approaches have not looked at co-design of multiple levels in the system software stack (OS, runtime, compiler, libraries, application frameworks)
- Need to exploit new hardware trends (e.g., manycore, heterogeneity) that cannot be handled by existing software stack, memory per socket trends
- Emerging software technologies exist, but have not been fully integrated with system software, e.g., UPC, Cilk, CUDA, HPCS
- Community codes unprepared for sea change in architectures
- No global evaluation of key missing components

IESP Goal

Improve the world's simulation and modeling capability by improving the coordination and development of the HPC software environment

Workshops:

Build an international plan for developing the next generation open source software for scientific high-performance computing

Where We Are Today:

- ☐ SC08 (Austin TX) meeting to generate interest — Nov 2008
- ☐ DOE's Office of Science funding
- ☐ US meeting April 6-8, 2009 — Apr 2009
  - ☐ 65 people
- ☐ NSF's Office of Cyberinfrastructure funding
- ☐ European meeting June 28-29, 2009 — Jun 2009
  - ☐ 70 people
  - ☐ Draft Roadmap
  - ☐ Outline Report
- ☐ Asian meeting (Tsukuba Japan) October 18-20, 2009
  - ☐ Refine roadmap
  - ☐ Refine Report
- ☐ SC09 (Portland OR) BOF to inform others — Oct 2009
  - ☐ Public Comment
  - ☐ Draft Report presented — Nov 2009

All of these issues add programming complication

- Assertion: data structure design and data motion minimization have
- more impact on performance than instruction ordering
  - But, these are both architecture specific!
- Resilience: DARPA exascale report has component failure 35-39 mins
  - Message delivery failure in MPI-3
  - Dead node detection and recovery
  - Needs to be integrated from the hardware through the application
- Soft error tolerance
  - If we assume any operation can give incorrect results, can we make more robust algorithms?
  - Can we better protect high-order bits?
- Some hardware support libraries are only available in certain programming languages, and some programming models only on certain hardware

Conclusions

- For the last decade or more, the research investment strategy has been overwhelmingly biased in favor of hardware.
- This strategy needs to be rebalanced - barriers to progress are increasingly on the software side.
- Moreover, the return on investment is more favorable to software.
  - Hardware has a half-life measured in years, while software has a half-life measured in decades.
- High Performance Ecosystem out of balance
  - Hardware, OS, Compilers, Software, Algorithms, Applications
    - No Moore's Law for software, algorithms and applications

Collaborators / Support

Employment opportunities for post-docs in the ICL group at Tennessee

- Top500
  - Hans Meuer, Prometeus
  - Erich Strohmaier, LBNL/NERSC
  - Horst Simon, LBNL/NERSC

If you are wondering what's beyond ExaFlops

Mega, Giga, Tera, Peta, Exa, Zetta ...

10^{3}	kilo
10^{6}	mega
10^{9}	giga
10^{12}	tera
10^{15}	peta
10^{18}	exa
10^{21}	zetta

10^{24}	yotta
10^{27}	xona
10^{30}	weka
10^{33}	vunda
10^{36}	uda
10^{39}	treda
10^{42}	sorta
10^{45}	rinta
10^{48}	quexa
10^{51}	pepta
10^{54}	ocha
10^{57}	nena
10^{60}	minga
10^{63}	luma

科	学	技	術	計	算	分	科	会		選	出

SS 研 HPC フォーラム 2009 より

JAXA Supercomputer System(JSS)の紹介と性能概要

宇宙航空研究開発機構

高木　亮治

サイエンティフィック・システム研究会 2009 年 9 月 3 日
SS 研 HPC フォーラム 2009

JAXA Supercomputer System (JSS) の紹介と性能概要

高木 亮治、藤田直行、松尾裕一
宇宙航空研究開発機構

[アブストラクト]

宇宙航空研究開発機構(JAXA)は航空宇宙分野における基礎研究から研究・利用までを一貫して行っており、前身の宇宙三機関の時代から高性能計算機を用いた数値シミュレーション技術の重要性を認識し、高性能・高機能な大規模計算機システムの整備・運用を積極的に推進してきた。2009 年 4 月に JAXA Supercomputer System(略して JSS)と呼ばれる新しいシステムが稼動を開始した。JSS は複数の計算機システムから構成されるが、その中核は富士通製 FX1 で、マルチコアスカラーCPU を用いた大規模超並列計算機であり、120TFlops の理論演算性能と 94TByte の主記憶容量を持っている。

本講演では JSS の概要を紹介すると同時に、JAXA で実際に使われている航空宇宙分野における CFD プログラムを用いた性能評価結果について報告する。

[キーワード]

航空宇宙、CFD(計算流体力学)、大規模並列計算機、マルチコア

1. はじめに

宇宙航空研究開発機構(JAXA)は、航空宇宙分野の基礎研究から開発・利用までを一貫して行っているが、前身の航空宇宙技術研究所(NAL)および宇宙科学研究所(ISAS)の時代から高性能計算機を用いた数値シミュレーション技術の重要性を認識し、高性能・高機能な大規模計算機システムの整備・運用を積極的に行ってきた。航空宇宙分野における数値シミュレーションは大規模な解析が多く、必然的に大規模並列計算機システムが必要とされてきた。単純な 2 次元翼型まわりの流れの解析から始まり、計算機の発達とともにより複雑な形状、例えば航空機全機まわりの粘性流れ解析や、より複雑な現象、エンジン内での化学反応を伴う燃焼流れなどの解析へと発展していった。これらの数値シミュレーションは学術研究のツールとしてだけではなく、実際の航空機、ロケット、衛星・探査機などの設計や開発に応用されている。特に実際の応用においては開発において発生したトラブルに対するトラブルシューティング的な課題解決型のアプローチから、徐々にではあるが設計探査や最適化といった設計プロセスを革新する様な使われ方にシフトしている。

JAXA 統合前後において、旧 3 機関時代からの経緯で統合後も調布、角田、相模原の 3 箇所で大規模計算機システムが運用されていたが、調布、角田のシステムがほぼ同時期にリースアウトするのを契機に 2009 年 4 月に新しい大規模並列計算機システムを導入した。新しく導入した計算機システムは JSS(JAXA Supercomputer System)と呼ばれ、JAXA 統合後初めての導入となることから、これまで以上に宇宙開発等の JAXA 事業への本格的な活用および宇宙三機関統合のシンボル的な位置づけ(One-JAXA)を意図して導入された。

本報告では、まず始めに JSS の設計思想およびシステム構成、特徴等について紹介する。次に JAXA で利用されている代表的な CFD プログラムを用いた JSS の性能評価結果について報告する。最後に JSS 導入時に実施された大規模解析について紹介する。

2. 設計思想

JSS を設計するにあたり様々な角度からの検討を行った。まず旧システムの課題として大規模 SMP の使い難さがあった。JAXA では SMP ノード内に複数のジョブが混在する運用を行っていたためジョブの計算時間のぶれが発生した。計算時間のぶれはプログラムチューニングの大きな障害となり最後まで抜本的な解決は行えなかった。またメモリバンド幅不足、自動スレッド並列コンパイラの能力不足、スカラーCPU の経験不足などから期待した性能が出せなかった。

本来、どのような計算機（演算性能重視、メモリ性能重視、通信性能重視）を導入すべきかは利用するアプリケーションに大きく依存する。最近の JAXA アプリの傾向として工学系アプリの Capacity 計算指向および学術系アプリの Capability 計算指向が挙げられる。工学系はパラメトリック計算などスループット重視のものであり、学術系は性能や規模が重視される。これらのアプリはさらに相対的に計算負荷の大きいアプリ（計算系）、通信負荷の大きいアプリ（通信系）、メモリアクセス負荷の大きなアプリ（メモリ系）に分類できる。どのような特性を持ったアプリをターゲットとするかで計算機のバランス（演算性能、メモリ性能、ネットワーク性能）が決められるが、JSS では工学系アプリを中心に考えつつ学術系アプリにも配慮する方針とし、そのため演算性能とメモリ性能を重視することとした。

技術的な観点からは、先進的過ぎて実績がなかったり、維持管理（手間、コスト）が大変な技術・システムの採用はやはり困難であり、将来動向は見据えつつも確実な技術や持続可能な技術を採用する必要がある。例えばノード形態としてはメモリ性能や電力・コストを考えると大規模共有メモリノードよりも有利な小規模ノードを選択した。また結合ネットワーク（インターコネクト）に関しては、伝統的なクロスバーネットワークは物量的に非現実的であり、ファットツリーなど実績のある多段結合網とした。

最後に統合スパコンとしての要求に応えるため、これまでのプログラムの継続性や遠隔地からの利便性に配慮した。そのため各拠点にはフロントエンド機能やファイルサーバ機能を有する遠隔利用システムや分散データ共有システムを配置した。またベクトルジョブへの配慮から小規模ベクトルシステムを導入した。さらに、前後処理や非並列ジョブ、市販アプリの動作プラットフォームを考えた場合、大きなメモリ空間を有する共有メモリシステムは魅力的であり、巨大なメモリを有する共有メモリシステムを別途用意することとした。

3. システム構成と特徴

様々な要求項目や検討の結果、JSS は図 1 および表 1 で示す様に大規模並列計算機システム、ストレージシステム、共有メモリシステム、遠隔利用システム、分散データ共有システムなど複数の計算機システムから構成される複合システムとなった。従来システムに比べて演算性能は約 15 倍、メモリ量では約 25 倍、ストレージ量では約 20 倍程度の性能を有する。JSS の中で実際の計算の中核となるシステムは大規模並列計算機システムであり、マルチコア CPU をベースにした富士通製 FX1 と呼ばれるスカラー超並列計算機で構成される。大規模並列計算機システムは 120TFlops の演算性能と 94TBytes の主記憶装置を有する M（メイン）システムと 15TFlops、6TBytes の P（プロジェクト）システムから構成される。これら二つのシステムでは Integrated Multicore Parallel ArChiTecture （IMPACT） と呼ばれるマルチコア CPU を効果的に利用する技術が採用されている。IMPACT を構成するコア間ハードウェアバリア、共有キャッシュ、自動スレッド並列コンパイラの連携により従来性能が出せなかった細粒度スレッド並列（内側ループでのスレッド並列化）でも十分な性能が期待できるようになった。そのため、ユーザーにはノード間をプロセス並列を用いて並列化し、ノード内は IMPACT の自動並列コンパイラによる自動スレッド並列にまかせるという並列化モデルを推奨している。

共有メモリシステムは A(アプリケーション)システムと呼ばれる 1TBytes の共有メモリを有する富士通製 SPARC Enterprise M9000 と V(ベクトル)システムと呼ばれる、4.8TFlops の演算性能と 3TBytes のメモリを有する NEC 製ベクトル計算機システム SX-9 からなる。これらのシステムは巨大な共有メモリを持つ利点を活かして、前後処理を含む非並列ジョブや市販アプリ、ベクトルジョブの実行に利用される。

ストレージシステムとしてはディスクが 1PByte、テープが 10PByte ありディスクへの総実行転送性能は 28GByte/s(ioperf により測定)となる。ディスクとテープは階層型ストレージ管理(HSM)を行っている。

JSS の主要部分は調布事業所に設置されるため、角田、筑波、相模原などの遠隔拠点には遠隔利用システム(L システム)を設置した。これは各拠点からの利便性を向上させるためのもので、ファイルサーバやフロントエンドの役割を果たす。また各拠点間でのファイル共有を実現するため J-SPACE と呼ばれる分散データ共有システムを構築した。調布システムと各拠点のシステムとはスパコンネットと呼ばれる高速ネットワークで接続されている。

JSS 導入にあたり、既存建屋には入りきらないため新建屋を建設した。新建屋では冷却効率の向上を目指し、排気拡散防止版、空調ダクトを設置するなどして暖気と冷気をできるだけ分離し、冷却効率が高まるような工夫を行った。また一般的な電力による空調の代わりにガスによる空調機の採用や遮音板の設置、室内設定温度の検討など環境に配慮したものとなっている。

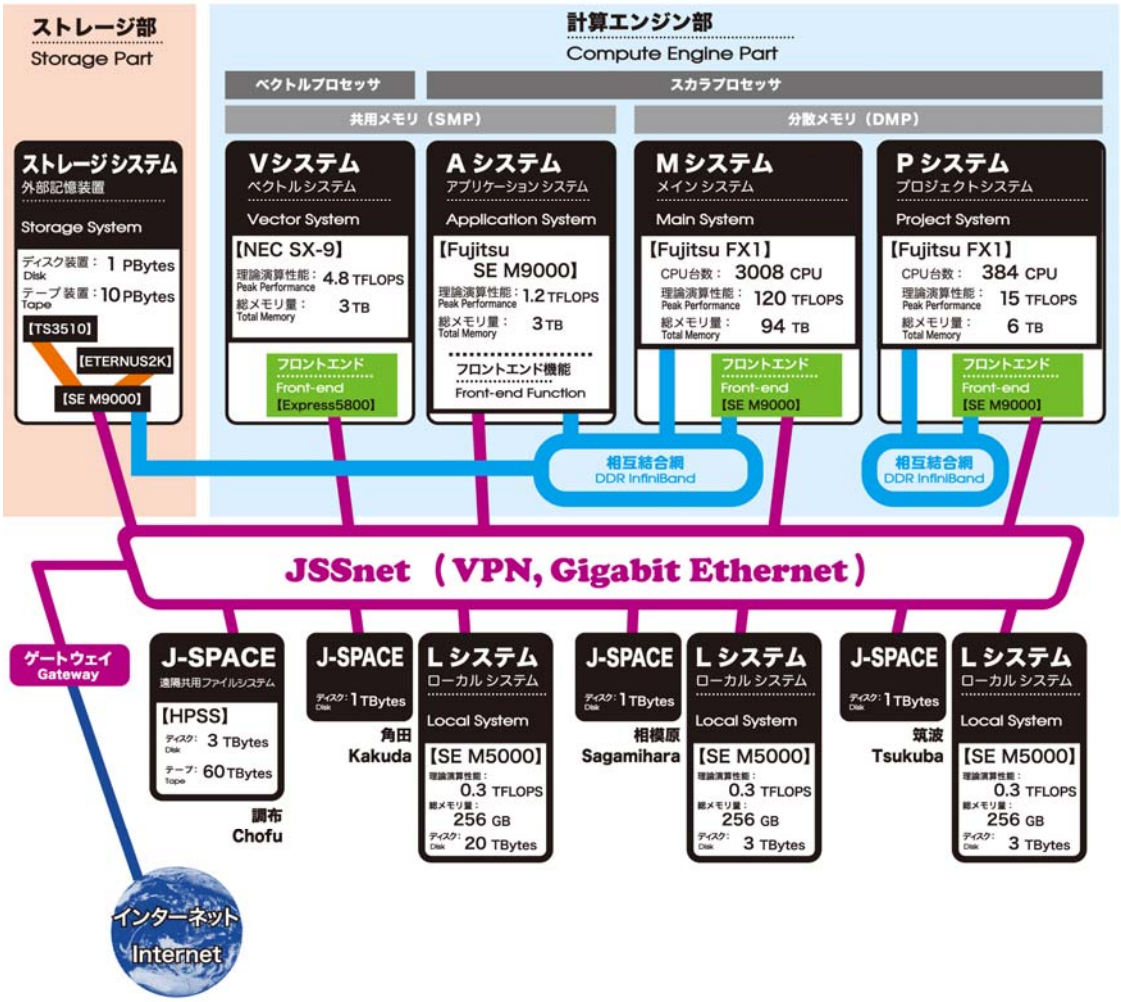


図 1 : JAXA Supercomputer System (JSS)のシステム図

表1:JSS の主要計算機システム

名称	JSS-M/P	JSS-A	JSS-V
CPU	Scalar	Scalar	Vector
System	MPP	SMP	SMP
ノード数	3008/384	1	3
CPU/ノード	1	32	16
Core/CPU	4	4	1
論理性能 [TFlops]	120/15	1.2	4.8
ノード性能 [GFlops]	40	40	1,600
総メモリ [TBytes]	94/6	1	3
ノードメモリ [GBytes]	32/16	1	1
メーカー	富士通 FX1	富士通 SEM9000	NEC SX-9

4. 性能評価

JSS における計算の中核となる JSS-M/V に関して実際に JAXA で使われている実アプリケーションによる性能評価を行った。まず JSS-M に対する評価について示す。表 2 にアプリケーションの概要を示す。どのアプリケーションも日常的に使われている実用アプリケーションであり、各アプリケーションの特性としては P1 と P4 は演算負荷が大きいアプリケーション、P2 とP5はメモリアクセス負荷が大きいアプリケーション、P3 はネットワーク負荷が大きいアプリケーションである。

表 2:JAXA アプリケーション一覧

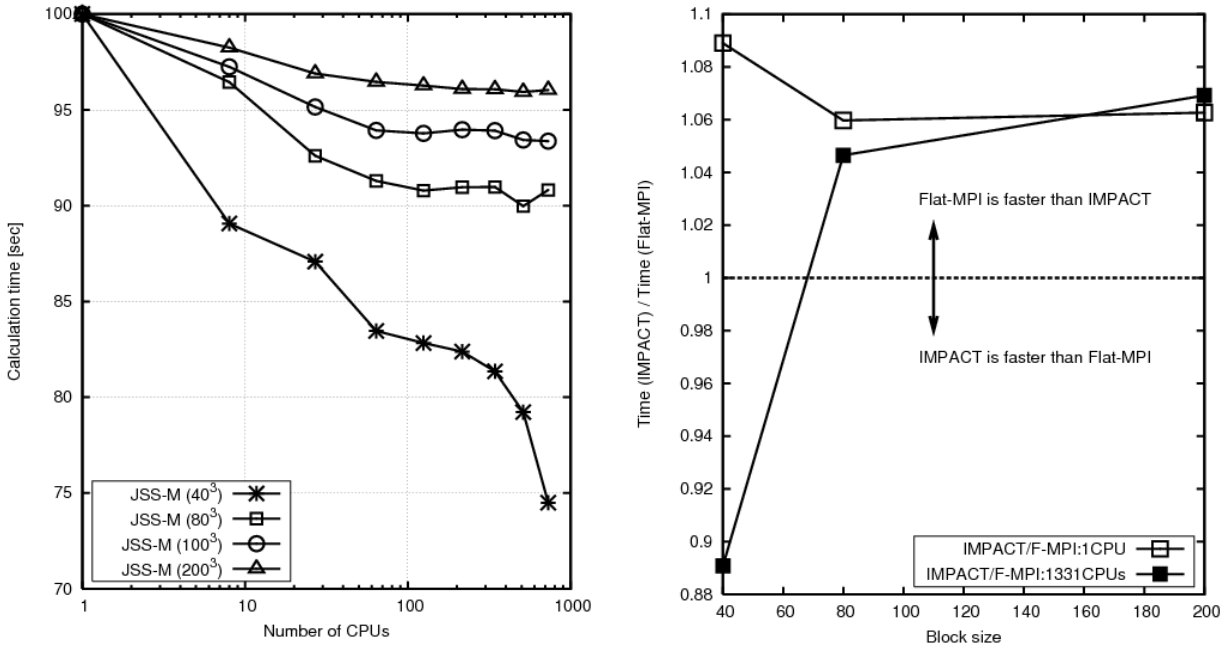
名称	適用先	計算手法	並列化	特性
P1	燃焼	FDM+化学反応	MPI+IMPACT	演算負荷が大
P2	航空	FVM (構造)	MPI+IMPACT	メモリアクセス負荷が大
P3	乱流	FDM+FFT	XPF+IMPACT	ネットワーク負荷が大
P4	プラズマ	PIC	MPI+IMPACT	演算負荷が大
P5	航空	FVM (非構造)	MPI+IMPACT	メモリアクセス負荷が大

表 3 に測定結果を示す。この結果より JAXA の実アプリケーションに対して JSS-M は CeNSS(従来システムである NS-III の中核計算機)より平均で 10 倍以上高速であることがわかる。P2 と P4 の性能比が他よりも高いが、P2 は自動スレッド並列コンパイラの性能向上によりスレッド並列の範囲が広がったためと考えられる。また P4 は MPI の集合通信の改善が主な理由と考えられる。逆に P3 が悪いのは他アプリケーションと異なり、演算負荷に対して相対的にデータ転送負荷が大きく経過時間ではほぼ同程度となる。JSS-M では通信性能比(対 CeNSS 比で 2 倍にしかならない)は演算性能比(周波数:2 倍、コア数:4 倍、その他:?倍)に比べて悪いため、全体の性能比が悪くなったと考えられる。

表 3:JAXA アプリによる性能評価

名称	CPU 数	CeNSS [sec]	JSS-M [sec]	性能比
P1	744	1380.4	143.3	9.63
P2	750	1468.6	91.5	16.05
P3	512	3517.0	491.7	7.15
P4	750	3061.7	193.0	15.86
P5	750	1447.2	181.6	8.13
平均 (相乗平均)	-			11.36 (10.73)

JAXA の代表的な実アプリの一つである UPACS を用いて JSS-M の性能評価を行った。UPACS は構造体、動的配列、ポインター、モジュールといった Fortran90 の機能を活用して作成された汎用的な圧縮性流体解析プログラムであり、3 次元マルチブロック構造格子および重合格子を扱うことができる。UPACS を用いたスケールアップ評価を図 2a)に示す。使用する CPU を増やす度に計算規模を増大させるスケールアップ評価では 729CPU まで 74%(ブロックサイズは 40^3)～96%(ブロックサイズは 200^3)といった良い並列効率を示した。ブロックサイズが大きくなると相対的に並列性能は良くなる。何故なら、ブロックの1辺をNとすると演算量は N^3 に対して通信量は N^2 に比例するため、N が大きくなる、つまりブロックサイズが大きくなると相対的に通信によるオーバーヘッドが減少するからである。



a) スケールアップ評価　　b) IMPACT とフラット MPI の比較

図2:UPACS による JSS-M の評価

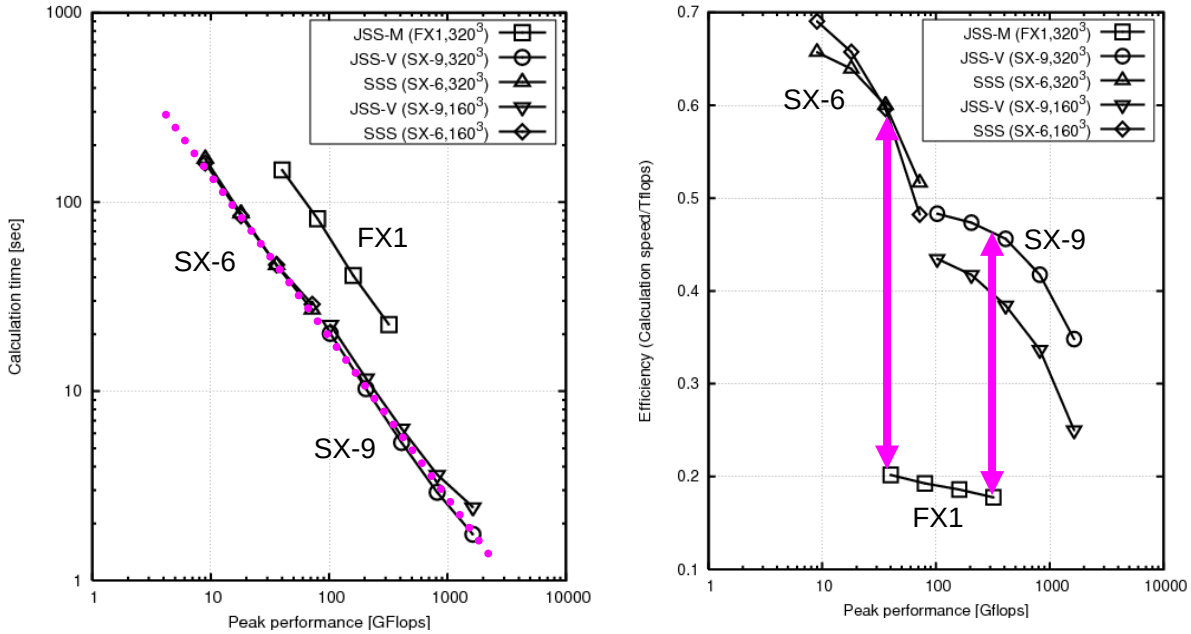
IMPACT とフラット MPI の比較を図 2b)に示す。IMPACT とフラット MPI で、どちらが良いかは色々な条件の影

響を受ける。例えば、CPU 内並列を考えた場合、IMPACT の並列性能は自動並列コンパイラの能力に依存するが、フラット MPI ではユーザーが明示的に並列化を行うことで高い並列性能が期待できる。一方、プロセス間通信を考えた場合、IMPACT はプロセス数を減少させることができるので、高プロセス並列においても性能劣化が低いが、フラット MPI の場合はプロセス数がコア数倍増えるため、高プロセス並列においては性能劣化が大きいと考えられる。またアプリ側の問題として計算格子のブロックサイズが大きくなるとフラット MPI が有利となる。これは前述したようにブロックサイズが大きくなると通信のオーバーヘッドが小さくなるからである。ブロックサイズの影響は図 2b)からも読み取れる。現状の JSS の規模ではアプリケーションにもよるが IMPACT よりもフラット MPI の方が若干有利かもしれないが、今後更なるスケールアップ（コア数、ノード数）を考えた場合、フラット MPI の限界が見えてきたと考えている。

次にこれも JAXA の代表的なアプリの一つである LANS3D を用いて JSS-M および JSS-V の評価を行った。LANS3D は航空宇宙分野で比較的初期に開発された先駆的な CFD プログラムで Fortran77 をベースに構造化プログラミング的な考え方で設計され、主にベクトル計算機向けに実装された典型的な圧縮性流体解析プログラムである。並列化に関しては基本的に自動スレッド並列による並列化を行っている。プロセス並列化としては MPI を用いた領域分割に一部対応しているが、多数プロセスによる並列計算は現実的でないためここでは最大 8 プロセスまでとした。この LANS3D を用いて JSS-M/V のスピードアップ評価を行った。問題規模として約 3300 万点の計算格子を固定して、CPU 数を増やすことによる計算速度の向上を評価した。表 4 に計算の概要を示す。JSS-M では IMPACT を用いて、プロセス数として1～8プロセスまで測定を行った。プロセス数に応じて計算格子をブロック分割し、1CPUに1ブロックを割り当てた。JSS-V および SSS（現在も相模原で運用中の NEC 製 SX-6）ではノード内に限定し自動並列でスレッド数1～16（SSS は8まで）で計測を行った。またベクトル長の影響を見るため、320x320x320 が1ブロックの計算と 160x160x160 が8ブロックの計算を行った。測定結果を図 3 に示す。図 3a) は計算時間を比較したもの、図 3b) は相対的実行効率を比較したものである。ここで相対的実行効率は計算時間の逆数を利用したCPUのピーク性能で割った値であり、単位ピーク性能あたりの計算速度を示す。そのためこれらの値の比較は実行効率の比較と同じ意味を持つ。相対的実行効率は通常の実行効率とは異なり、利用範囲が限定されるが、ユーザーの実感に近い、計測が容易といったメリットがあるため、ここでは相対的実行効率を比較した。図 3b) より SSS(SX-6)と JSS-M(FX1)では実行効率で 3 倍程度の違いが見られるが、JS-V(SX-9)と JSS-M(FX1)では約 2.5 倍程度の違いに縮小していることがわかる。また JSS-V(SX-9)は SSS(SX-6)に比べてベクトル長が短くなると性能が悪化することもわかる。

表 4：LANS3D による評価

システム	プロセス数	スレッド数	ピーク性能 [GFlops]	(ブロックサイズ) ×ブロック数	並列手法
JSS-M	1	4	40	(320x320x320) x1	IMPACT
	2	8	80	(320x320x160) x2	MPI+IMPACT
	4	16	160	(320x160x160) x4	
	8	32	320	(160x160x160) x8	
JSS-V(SX-9)	1	1,2,4,8,16	102.4～1638.4	(320x320x320) x1 (160x160x160) x8	自動並列
SSS(SX-6)	1	1,2,4,8	9.0～720	(320x320x320) x1 (160x160x160) x8	



a) 計算時間の比較　　b) 相対的実行効率の比較

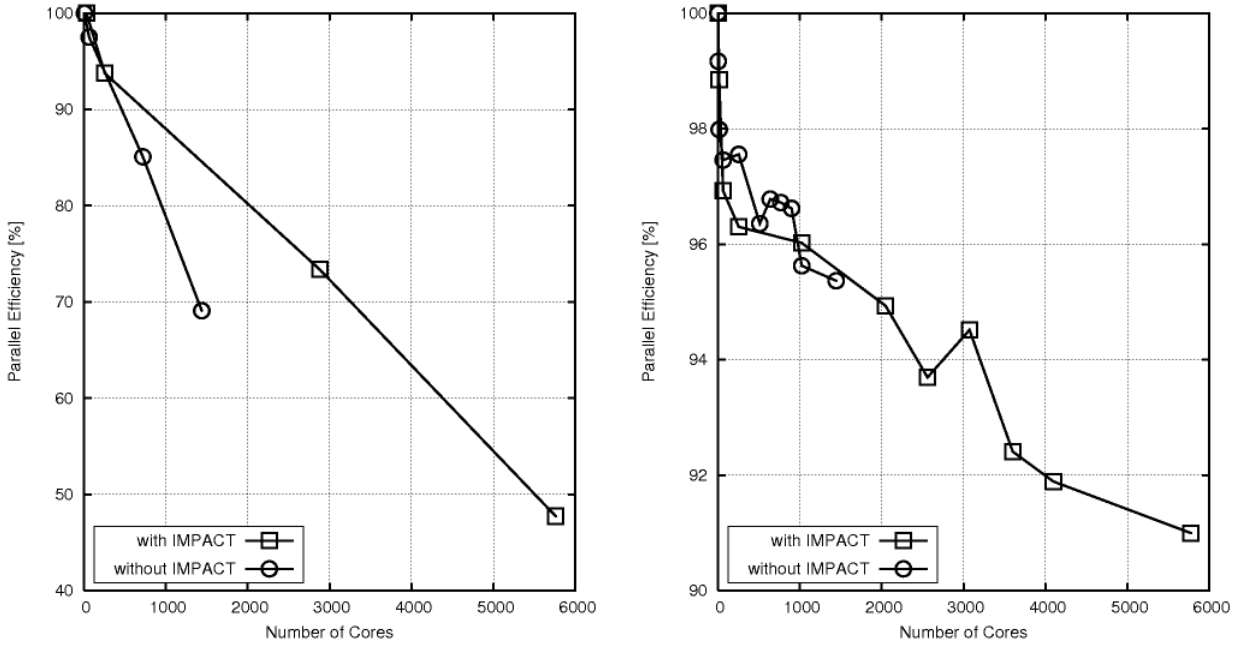
図 3:LANS3D による評価結果(JSS-M(FX1), JSS-V(SX-9), SSS(SX-6))

5. 大規模解析

JSS 導入時に JSS-M の半分を1ヶ月程度利用する大規模解析を実施した。「液体燃料噴流微粒化過程解明の大規模計算(LSC1)」と「大規模粒子計算で探る宇宙空間衝撃波のダイナミクス ～科学衛星観測成果の理解に向けて～(LSC2)」の 2 件でそれぞれの計算概要を表 5 に、スケールアップでの並列化効率を図4に、計算結果の一例を図5に示す。LSC1 は計算時間ネック、LSC2 はメモリネックの解析である。また並列化効率も LSC1 では 50%弱であるが、LSC2 は 90%以上の高い値となった。それぞれの計算結果は当該分野で大きな成果を挙げているが、それらの成果もさることながら、導入初期に大きなトラブルもなくこれらの計算が延べ2ヶ月に渡って実行できたことはシステムの安定性・信頼性が非常に高いことの表れである。これとは別に一般的な Linpack HPL の計測も行ったが、その際も延べ 60 時間以上の高負荷計算においてもトラブルなく計算を完遂することができ、システムの安定性が非常に高いことを実証できた。なお、Linpack に関しては Rpeak=121.282TFlops に対して Rmax=110.600 となり高い実行効率(91.19%)を示した。

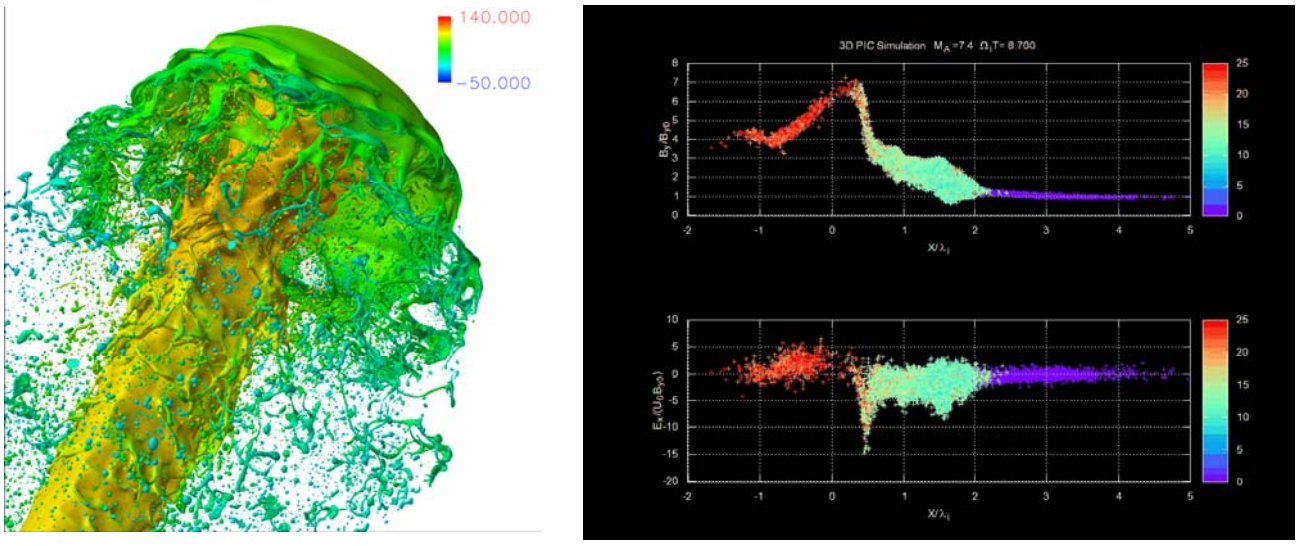
表 5:大規模解析の概要

	LSC1	LSC2
並列規模	1440 プロセス(5760 コア)	1444 プロセス(5776 コア)
計算規模	格子点:58 億点	格子点:4.5 億点、粒子数:500 億個、メモリ:40TB
計算時間	410 時間	740 時間
出力ファイル	153TB(25 時間)	180TB(総量:430TB、43 時間)
実行効率	約 4%程度	約 8%程度



a) LSC1　　b) LSC2

図4：大規模解析の並列化効率（スケールアップ）



a) LSC1 の結果　　b) LSC2 の結果

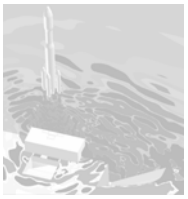
図5：大規模解析の計算結果例

6.　おわりに

JAXA の新スパコンシステム（JSS）について、設計思想、システムの概要、初期性能評価、大規模解析例などについて紹介した。今後は更なる性能評価、ユーザーのチューニング支援、可視化システムの整備などを進めて行くと同時に、JSS を活用した数値シミュレーション技術により、JAXA 事業における設計開発プロセス自体の革新を目指し、「もの造り」への貢献をより強力に推進して行く。

以上

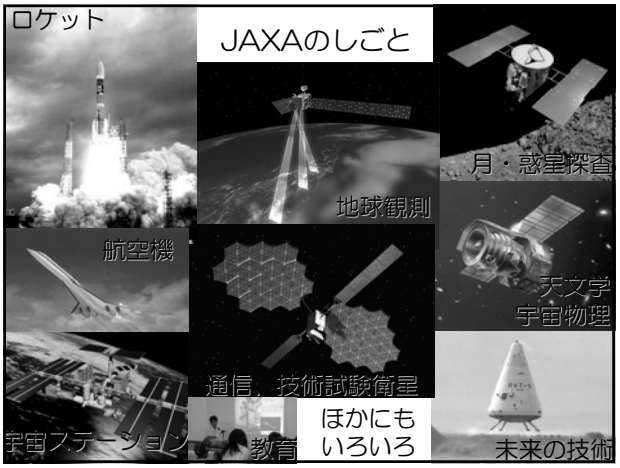
JAXA Supercomputer System(JSS)の紹介と性能概要



高木亮治，藤田直行，松尾裕一
（宇宙航空研究開発機構）

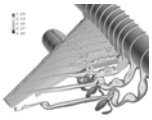
内　容

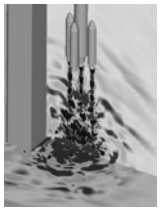
- 背景
- JSSの設計思想
- システム構成と特徴
- 性能評価
- 大規模解析
- まとめ



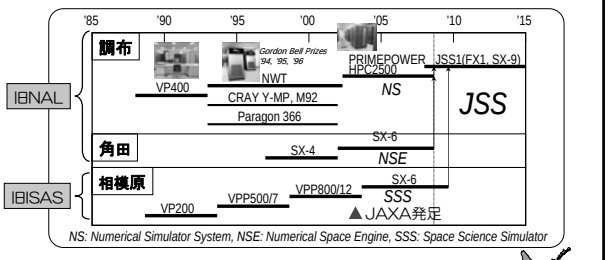
JAXAにおける数値シミュレーション技術

- 学術研究のツール
  - 宇宙科学を中心に
- 航空機、ロケット、衛星・探査機の設計・開発
  - 信頼性向上、開発期間の短縮、コスト削減、先進的技術の開発…
  - 数値シミュレーション技術の活用を重点化
    - **基礎実験/データ、打ち上げ実績：欧米と大きな差**
- アプローチの仕方
  - 課題解決
    - 現象理解（極限状態）
  - 設計プロセスの革新
    - 概念検討、最適化（設計探査）





JAXA統合前夜



NS: Numerical Simulator System, NSE: Numerical Space Engine, SSS: Space Science Simulator

システム名	NS	NSE	SSS
ピークTFLOPS値	9.3	0.5	1.1
主記憶容量[TByte]	3.7	0.5	1
総CPU数	1792	64	128
演算器の種類	スカラ	ベクトル	ベクトル

JSSの設計思想

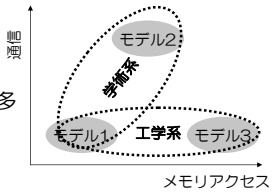
旧システム（主にNS）の課題

大規模SMPノードの課題

- 利用者側から、
  - 資源枯渇
  - ベクトルからスカラーへの転換
  - 性能が出ない
    - メモリバンド幅の不足、自動並列コンパイラの能力不足
  - 大規模スレッドの使いにくさ
    - 実行時間のブレ → チューニングができない
- 運用者側から、
  - サイトが別々である事の弊害
    - 運用部隊の非効率性、情報の一元管理、利用技術の共用
    - JAXA統合効果
  - セキュリティレベル
  - 高額なソフトウェア、ライセンス管理

アプリケーションからの視点

- アプリモデル(1)
  - 工学系　⇒　スループット重視（Capacity計算指向）
  - 学術系　⇒　性能、規模重視（Capability計算指向）
- アプリモデル(2)
  - モデル1：計算多
  - モデル2：通信多
  - モデル3：メモリアクセス多



どこを中心に設計するか？

アプリケーションからの視点

- 基本要件
  - 性能向上（目標：150TFlops）
  - メモリ性能（B/F=1）
  - 工学系中心、学術系にも配慮
- システム要件
  - ノード性能、メモリ性能、メモリ量は高い・多い方が良い
  - ノードは，SMP等の大規模ノードでなくて良い
  - 現状の並列モデルを大きく変えない
  - ノード内並列は自動化
  - 通信はある程度の範囲で速く

運用サイドからの視点

- 基本要件
  - 安定稼動（トラブルが少ない、設定が楽）
  - 運用管理が楽、運用コスト少
  - 設置性（スペース、電力、冷却）
  - ユーザに対して同一サービスを提供
    - 遠隔からの利用に対して
  - ソフトウェアの移植性、汎用性
  - システムの拡張性
  - 性能情報が取得可能なこと

技術トレンドからの視点

- 基本技術
  - 先端的過ぎる、実績がないものは危険
    - 使いこなすのが大変
  - 5年間分の需要と将来動向を見据えた確実な技術
    - 当面マルチコア
- ノード
  - 小規模ノードが有利（メモリ性能、電力、コスト）
- 結合ネットワーク（インターコネクト）
  - クロスバは無理
  - フルバンドの多段結合網

検討結果

- CPUまわり
  - プロセスあたりの性能はできるだけ高くしたい
    - 流体解析は計算量が多い
  - 計算ノードに大規模SMPは性能面で不利
    - 電力、コスト、メモリ性能
  - スカラーCPUの場合、マルチコアを如何に有効に利用するかが課題
- メモリまわり
  - 運用実績より、RAM比（TB/TF）=0.6～0.8（90～120TB）で十分
  - ノードのメモリとして、数10GBは確保したい
  - 後処理や非並列ジョブのために、ある程度の規模の共有メモリ（数100GB）ノードが別にあると便利
  - メモリバンド幅は、B/F比=1 程度以上は必要
- 結合NWまわり
  - 高速な通信はこの範囲（全体の1/4まで）で行われれば良い。
    - 全システムを使うジョブはなく、最大でも1/3システム程度。通常は1/20～1/4が多い。

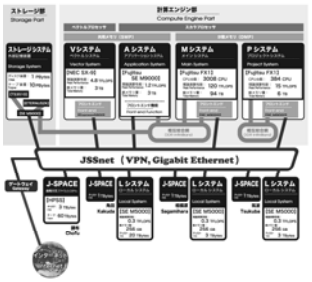
システム構成と特徴

システム構成

計算エンジン部
- 大規模並列計算機システム
- 共有メモリ計算機システム

ストレージ部
- ストレージシステム

分散環境統合部
- 遠隔利用システム
- 分散データ共有システム
- 高速ネットワーク



概要

- 国内最高クラスの性能
  - スカラー：135TFLOPS、ベクトル：4.8TFLOPS
- 世界最高クラスのLINPACK実行性能：91.19%
- 実用計算志向、使い勝手・円滑な移行に配慮
  - 複数のアーキテクチャが混在：選択の自由
  - 大規模メモリ：100TB以上
  - 大規模ストレージ：ディスク 1PB、テープ10PB
  - 共有メモリシステム：1TB共有メモリ
- 遠隔地からの利用環境
  - JSSネット：SINET3、VPNによる高速接続
  - ローカルシステム

計算エンジン部

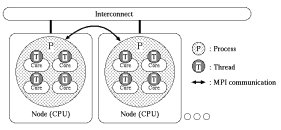
システム名称	Mシステム	Pシステム	Aシステム	Vシステム
CPUタイプ	スカラー			ベクトル
システムタイプ	MPP		SMP	
ノード数	3008	384	1	3
CPU数/ノード	1		32	16
コア数/CPU（全コア数）	4（12,032）	4（1536）	4（128）	1（48）
ピーク性能[TFlops]（ノードあたり[GFlops]）	120（40）	15（40）	1.2（40）	4.8（1600）
メモリ容量[TByte]（ノードあたり [GByte]）	94（32）	6（16）	1（1000）	3（1000）
製品名	富士通 FX1		富士通 SEM9000	NEC SX-9

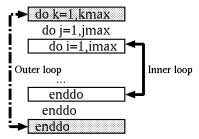
計算エンジン部（JSS-M/P）

- 富士通製FX1クラスタ
- 3,008ノード（12,032コア）
  - SPARC64TM VII 2.5GHz、4コア
  - 40GFlops、32GByte@ノード
- 94ラック
  - 32ノード、12KW@ラック
- FBBファットツリー・インターコネクト
  - DDR Infiniband
- 120TFLOPS、94TB
  - Linpack：110.6TFLOPS、91.19%
- JSS-Pはサブセット（15TFLOPS）

計算エンジン部（JSS-M/P）

- Integrated Multicore Parallel ArChiTecture：IMPACT
  - コア間ハードウェアバリア
  - 6MB 共有L2キャッシュ
  - 自動スレッド並列コンパイラ
    - 最内ループ並列でも性能が出る
- 高メモリ性能
  - 高メモリバンド幅 40GB/s
  - 低レイテンシ
  - 高信頼性（チップキル ECC）
- ノード間高速バリアネットワーク
  - データ転送とは別
  - ノード間ハードバリア
  - 集合通信のハードウエアサポート
  - OS割り込みによる遅延低減





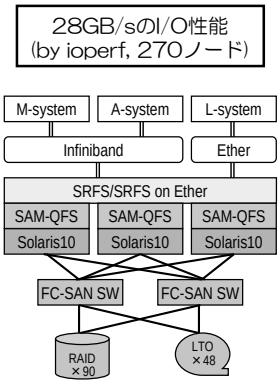
計算エンジン部（JSS-A/V）

- 1TBの大規模共有メモリマシン
- 前後処理、非並列ジョブ、特殊ジョブ（ベクトル向けジョブ、市販アプリ）

- JSS-A：富士通製SEM9000、1ノード
  - SPARC64TM VII 2.5GHz、4コア、40GF@ﾁｯﾌﾟ
  - 32CPU(128コア）、1.2TFLOPS
  - Fluent, NASTRAN, FIELDVIEW
- JSS-V：NEC製SX-9、3ノード
  - 102.4GFLOPS@CPU
  - 16CPU、1.6TFLOPS ⇒ 4.8TFLOPS、3TB
  - ノード間はIXSで接続
  - ベクトル向けジョブ

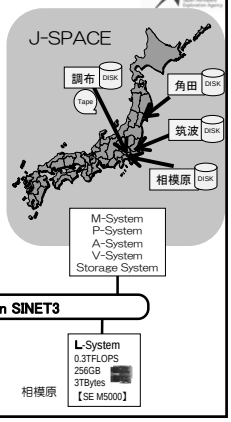
ストレージ部

- DISK：1PByte
  - RAID5
  - 4Gbps FC：180本
  - キャッシュ：360GB
  - SATA：7200rpm、750GB
- テープ：10PByte
  - 40×LTO4ドライブ、8×LTO3ドライブ
  - 4×TS3500ライブラリ、13,332カートリッジ
- I/Oサーバ：
  - SEM9000×3
- HSM：
  - SAM-QFS



分散環境統合部

- 遠隔利用システム（Lシステム）：主要事業所へのフロントエンド機能の提供
  - 角田、筑波、相模原
- インターネット(SINET3)越しの高速なファイル共有
  - SRFS on Ether
- 各拠点間でのデータ共有が可能な分散データ共有システム
  - J-SPACE(HPSS)



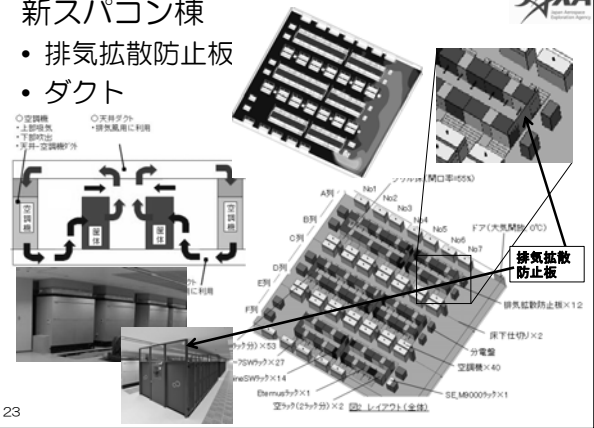
新スパコン棟

- 冷却効率の向上
  - 排気拡散防止板、空調ダクトで暖気と冷気を分離
  - 電力消費量の試算
  - ガス空調機
  - 防音対策
- 設備制御システム
  - 自動運転の実現
    - 起動・停止時
    - 負荷に応じた空調制御

空調機 吹出温度 [℃]	空調機 消費電力比 A	計算機 消費電力比 B	システム全体 消費電力比 C=A+B
20	1（基準）	2．00	3．00
25	0．95	2．08	3．03

新スパコン棟

- 排気拡散防止板
- ダクト



JSS-M（FX1）の性能評価

Linpack
JAXAベンチマーク
UPACS

Linpack HPL

- Top500のランキングに使われるベンチマーク
  - 高い実行効率：91.19%
  - 長時間安定稼動：60時間40分 → 耐久試験

Top500ランキング（国内分）

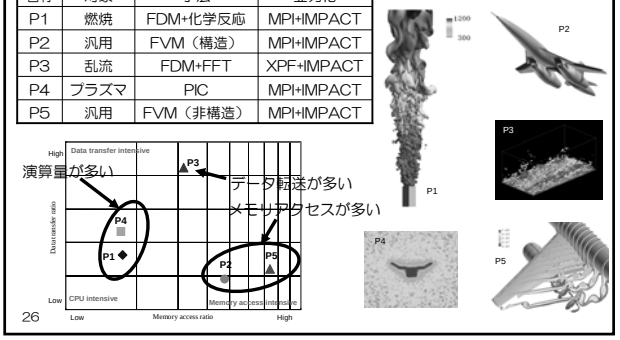
順位	サイト	マシン	コア数	Rpeak [TFlops]	Rmax [TFlops]	効率 [%]
22	地球シミュレータ	SX-9/E	1,280	131,072	122,400	93.38
28	JAXA	FX1	12,032	121,282	110,600	91.19
40	理研	RX200S5	8,256	96,760	87,890	90.83
41	東工大	Sun Fire	31,024	163,188	87,010	53.32
42	東大	HA8000	12,288	113,050	82,984	73.40
47	筑波大	Xtreme-X3	10,368	95,385	77,280	81.02

[2009.6現在]

JAXAベンチマーク

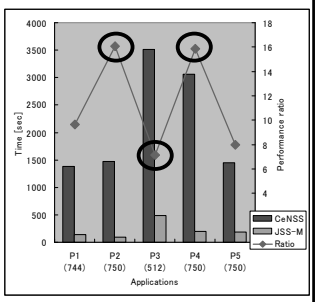
- JAXAの代表的なアプリケーション

名称	対象	手法	並列化
P1	燃焼	FDM+化学反応	MPI+IMPACT
P2	汎用	FVM（構造）	MPI+IMPACT
P3	乱流	FDM+FFT	XPF+IMPACT
P4	プラズマ	PIC	MPI+IMPACT
P5	汎用	FVM（非構造）	MPI+IMPACT



JAXAベンチマーク

- JSS-MIはCeNSS(旧ｼｽﾃﾑ)に比べて11倍(平均)高速
- CeNSS性能比
  - 演算性能：(8＋α)倍
    - クロック：2倍
    - コア数：4倍
    - その他：？
  - ネットワーク：2倍
- P2：自動並列コンパイラの改善
- P3：通信性能不足
- P4：集合通信の改善

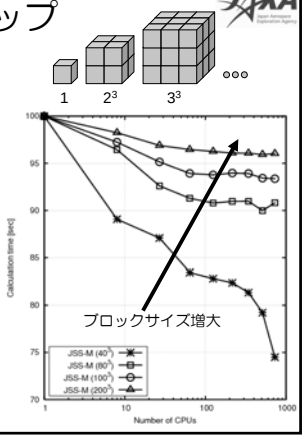


JAXAベンチマーク

			HPC2500 [sec]	FX1 [sec]	比	全体比
P1	演算		1373.0	138.4	9.9	9.63
	通信	隣接	7.40	4.6	1.5	
P2	演算		1465.60	79.3	18.5	16.48
	通信	隣接	1.48	3.0	0.5	
		リダクション	1.48	0.3	4.9	
		バリア	0.00	6.5	0.0	
P3	演算		1650.0	134.6	12.3	7.15
	通信	隣接	472.9	162.3	2.9	
		リダクション	1394.1	194.8	7.2	
P4	演算		2504.80	175.70	14.3	15.86
	通信	隣接	102.70	13.50	7.6	
		リダクション	616.30	3.60	171.2	
		バリア	0.02	0.20	0.1	
P5	演算		1208.6	150.00	8.1	8.13
	通信	隣接	125.2	10.20	12.3	
		リダクション	30.6	0.75	40.8	
		バリア	124.4	20.61	6.0	

3次元スケールアップ

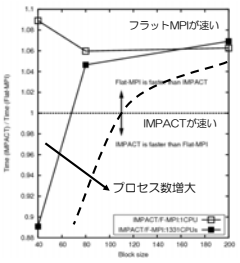
- UPACS
  - MPI+IMPACT
- ブロックサイズは40～200
- 729(9³)CPUまで良い並列効率(74～96%)を示す
- ブロックサイズが大きくなると並列効率が向上
  - 計算負荷(N³)と通信負荷(N²)のバランス



IMPACT 対 フラットMPI

UPACSを利用

- CPU内並列性能
  - 自動並列：コンパイラ
  - MPI並列：ユーザー指定
- プロセス間通信性能
  - プロセス数：小
  - プロセス数：大
- ブロックサイズ
  - ブロックが大：フラットMPI有利



	CPU数	プロセス数	スレッド数	ブロック数
IMACT:1CPU	1	1	4	1
フラットMPI:1CPU	1	4	1	4
IMPACT:1331CPUs	1331	1331	4	1331
フラットMPI:1331CPUs	1331	5324	1	5324

JSS-V（SX-9）の性能評価

LANS3Dによる評価

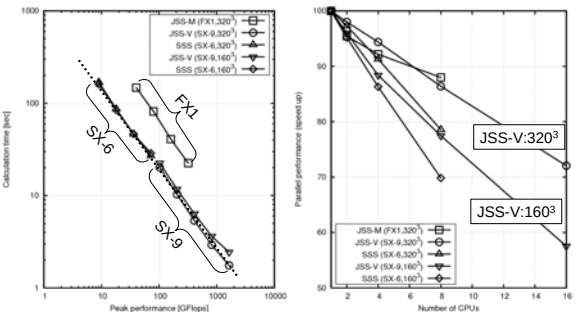
3次元スピードアップ

- 問題規模（約3300万点格子）を一定にして使用するCPU数を増やして速度向上性能を評価

システム	プロセス数	スレッド数	ピーク性能 [GFlops]	ブロックサイズ ×ブロック数	並列
JSS-M	1	4	40	(320x320x320)x1	IMPACT
	2	8	80	(320x320x160)x2	MPI+IMPACT
	4	16	160	(320x160x160)x4	
	8	32	320	(160x160x160)x8	
JSS-V	1	1, 2, 4, 8, 16	102.4～1638.4	(320x320x320)x1 (160x160x160)x8	自動並列
SSS	1	1, 2, 4, 8	90～720	(320x320x320)x1 (160x160x160)x8	

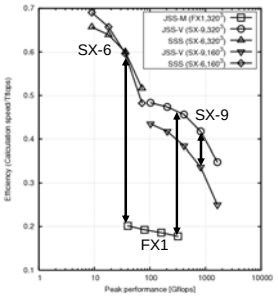
3次元スピードアップ

- 計算時間（左図）と並列効率（右図）



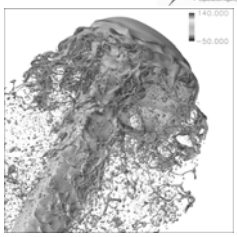
3次元スピードアップ

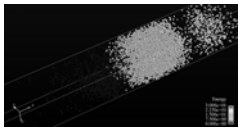
- 相対的実行効率
  = 1/計算時間/ピーク性能
  - 利用範囲は限定的
  - ユーザーの実感に近い
  - 計測が容易
- FX1に対して
  - SX-6：約3倍
  - SX-9：約2.5倍
- SX-9はベクトル長の影響が大



大規模解析の実施

- LSC1：液体燃料噴流微粒化過程解明の大規模計算
  - 新城淳史（JAXA/研究開発本部）
- LSC2：大規模粒子計算で探る宇宙空間衝撃波のダイナミクス～科学衛星観測成果の理解に向けて～
  - 篠原　育（JAXA/宇宙科学研究本部）



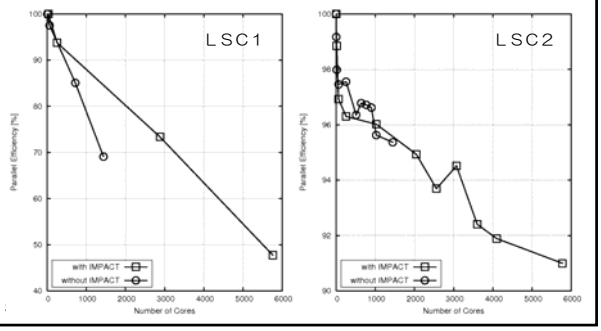


大規模解析

- LSC1
  - 並列規模：1440プロセス×4スレッド=5760コア
  - 計算規模：　（←計算時間ネック）
    - 格子点数：58億点
  - 計算時間：410時間
  - 出力ファイル：153TB（25時間）
  - 実効効率：約4%程度
- LSC2
  - 並列規模：1444プロセス×4スレッド=5776コア
  - 計算規模：　（←メモリネック）
    - 格子点数：4.5億点、粒子数：500億個、メモリ：40TByte
  - 計算時間：740時間
  - 出力ファイル：180TB（総量：430TB、43時間）
  - 実効効率：約8%程度
- 導入初期での安定稼動を実証

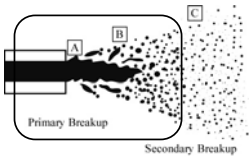
大規模解析

- 並列化効率（スケールアップ）
  – LSC1：50%弱、LSC2：90%強

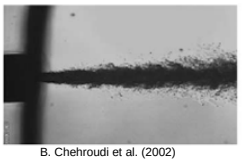


ＬＳＣ１：計算の目的と概要

- エンジン燃焼性能に重要な噴霧の物理を解明する
  - これまで噴霧の形成過程は謎のまま
  - 試行錯誤による非効率性と解析精度の悪さが問題



本計算で、噴霧の形成過程を根元から解明する（赤枠内）



B. Chehroudi et al. (2002)

実験では高速・小スケール・3次元のため流れ場を細かく見るのはかなり困難

- 気液二相流コード（表面張力あり）で最小液滴形状も解像できる格子で計算し、微粒化物理を解明する
- 総格子数約60億点と世界最大規模

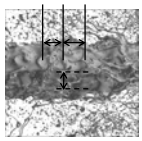
38

ＬＳＣ１：成果

– 微粒化へ至る詳細な物理を直接観察した
  - 特に液糸形成、液滴形成における表面張力波の役割について従来の考え方の不備を指摘。これにより従来の噴霧モデルを修正する作業へ移行中。

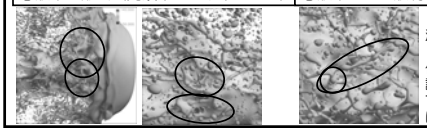
①縦方向と横方向の不安定

先頭傘による速度場乱れ



表面不安定の発達
（実線：縦方向、点線：横方向）

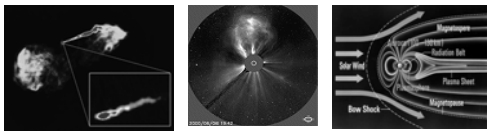
②液糸の形成（先頭傘および中心コア）　③液糸からの液滴の形成



液糸表面での表面張力「波」の形成とその不安定化による液滴生成物理を確認した。また、実際の噴霧では純粋なレーリーモードは起こりにくいことを指摘。

ＬＳＣ２：計算の目的と概要

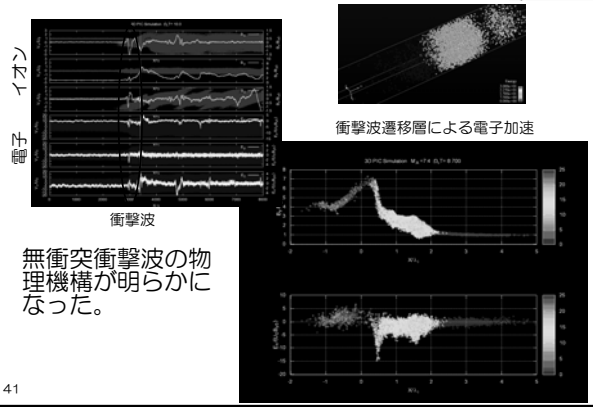
- 宇宙における衝撃波は高エネルギー粒子の生成現場→粒子加速の物理機構は何か？



- 科学衛星によりその場の観測が可能だが
  – 衛星観測の限界：絶対的な情報量の不足
  – 観測データの高精度化：データの解釈が難解
- 世界初の本格的な３次元計算により無衝突衝撃波の物理機構の解明、今後の科学衛星観測ミッションと連携

40

ＬＳＣ２：成果



無衝突衝撃波の物理機構が明らかになった。

41

JSS利用の今後

まとめ

- JAXAの新スパコンシステム（JSS）について紹介した。
  - 設計思想
  - システムの概要
  - 初期性能評価
  - 初期導入時に実施した大規模解析
- 利用の今後の方向性
  - 「もの造り」への貢献
  - 課題解決 → 設計開発プロセス自体の革新

今後の課題

- 詳細な性能評価→次期システムに向けて
- ユーザーのチューニング支援
- 大規模解析の継続←大規模システムの存在意義
  - 1,000ノード程度を定常的に
- 可視化システムの構築
  - 大規模可視化（並列可視化）
    - 例）現状：25GB → JSSでは500GB
      - 定常解析から非定常解析へ → 桁以上でデータが増大
  - 遠隔可視化
    - 例）相模原⇔調布間：500GBを転送すると7時間

可視化システムの試作