Opportunities & Challenges at Exascale

A Computational Science & Engineering Perspective

Thom H. Dunning, Jr.

Northwest Institute for Advanced Computing
Pacific Northwest National Laboratory & University of Washington

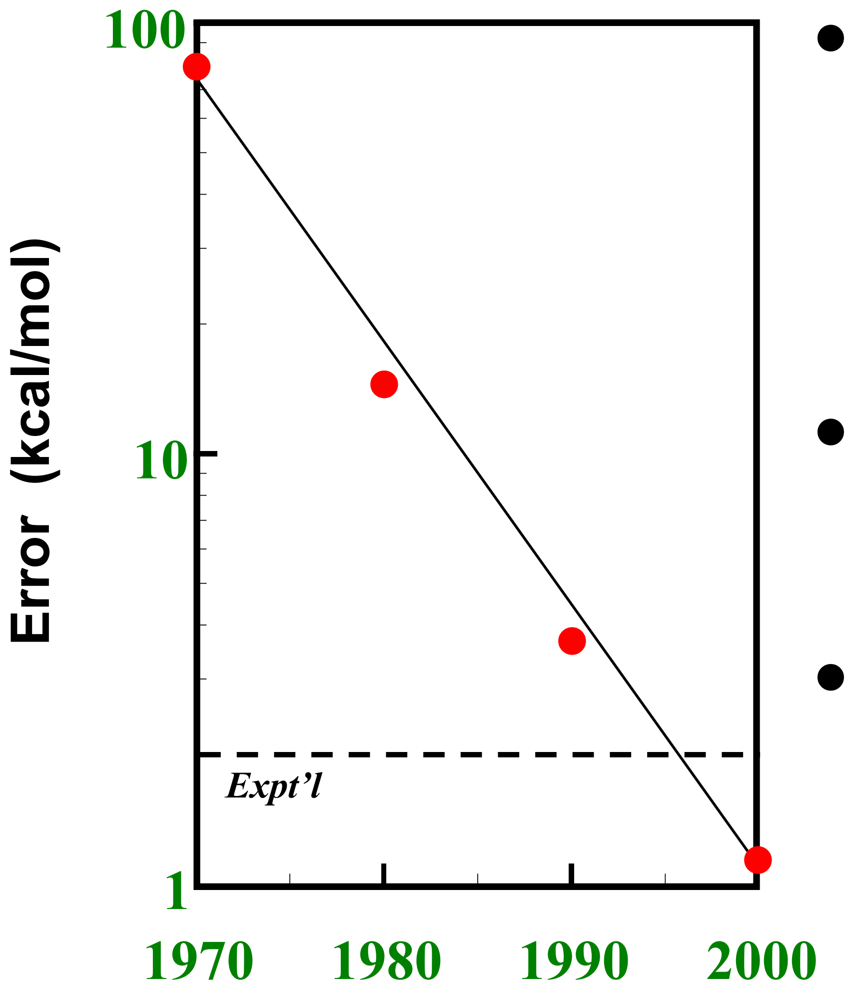
Department of Chemistry
University of Washington

Ever More & More FLOPS & Bytes

Computational scientists always seem to need more and more computing power and storage. What is the outcome of access to increasing amounts of flops & bytes?

More & More FLOPS & Bytes

Increasing Accuracy of Molecular Predictions

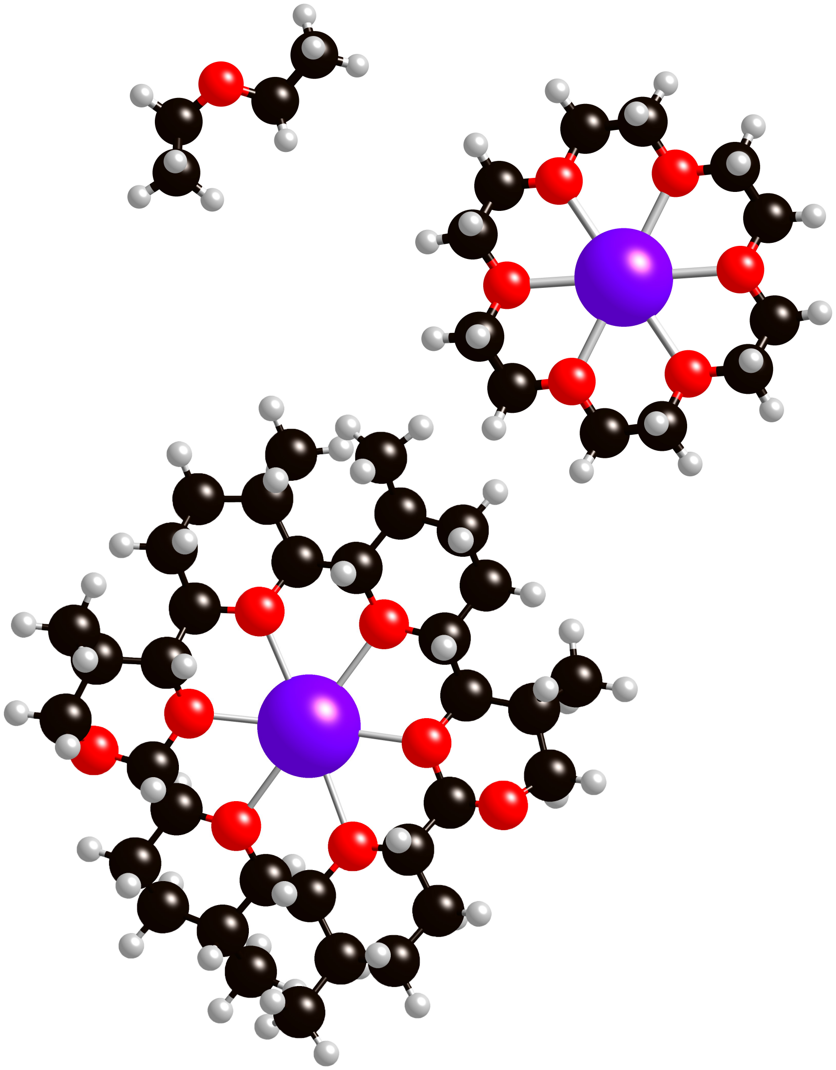


- **Bond Energies**
    - Critical for describing many chemical phenomena
    - Difficult to determine experimentally
- **Accuracy of Predictions**
    - Increased dramatically from 1970-2000
- **How?**
    - New theoretical approaches
    - New mathematical techniques
    - More computing power

More & More FLOPS & Bytes

Increasing Reach of Molecular Simulations

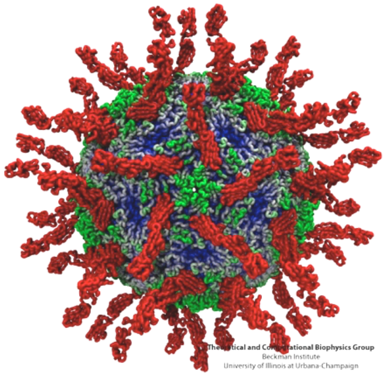
- **In 1990**
  - Model systems, e.g., ether–alkali ion complexes
- **In 2000**
  - Model separations agents, e.g., 18-crown-6–alkali ion complexes
- **In 2010**
  - Real-world separations agents, e.g., Still’s crown ether–ion complexes

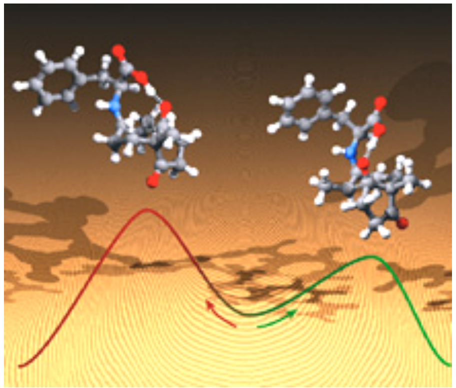


More & More FLOPS & Bytes

Similar Advances in Many Other Fields

Biomolecular Science

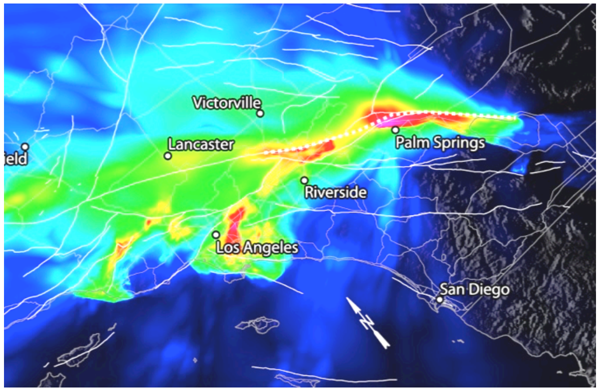




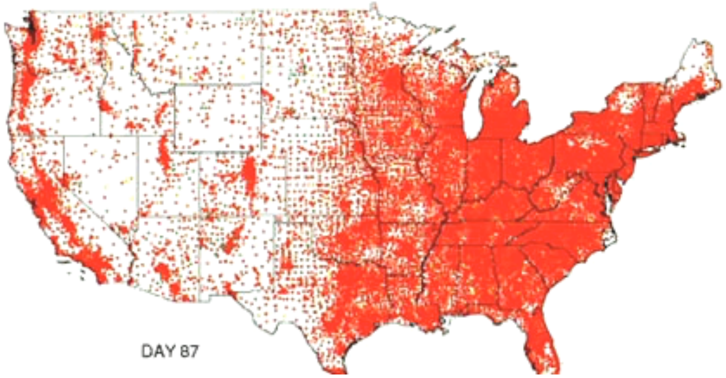
Weather & Climate

Astronomy

Geosciences



Health



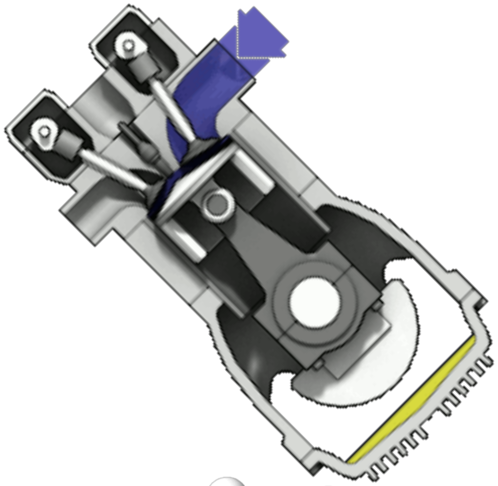
Petaflops & Petabytes

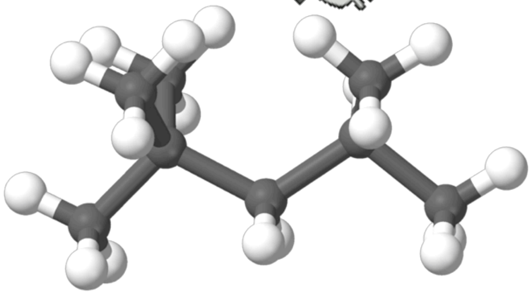
We are now in the petascale computing era. With these extraordinary computing capabilities scientists are further improving the fidelity of their models and increasing the complexity of the systems that they can model. Plus, entirely new applications are being explored.

Petaflops & Petabytes

Who Needs Petaflops?

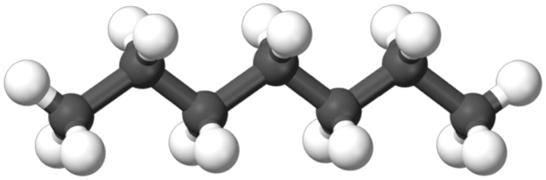
- **Energy content of Iso-octane**
  - Iterative solution of **275 million coupled equations**
  - Exchange of **2.5 petabytes of data** between processors
  - Exchange of **15 terabytes of data** between memory and disks
  - Execution of **30 quadrillion arithmetic operations**
- **Modeling Reactions of Fuels**
  - **Required** to understand **combustion** of fuels in engines





Iso-octane

(Octane Rating = 100)

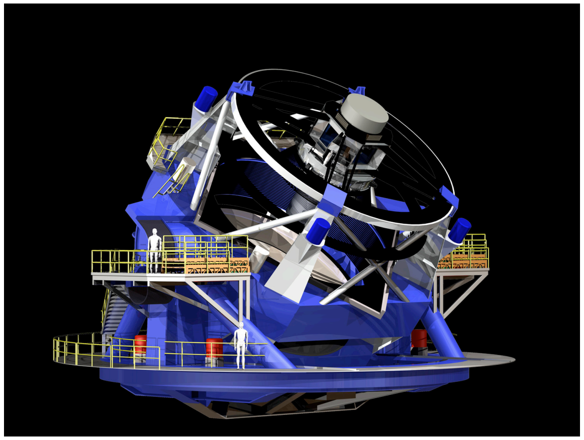


***n*-heptane**

(Octane Rating = 0)

Petaflops & Petabytes

Who Needs Petabytes?

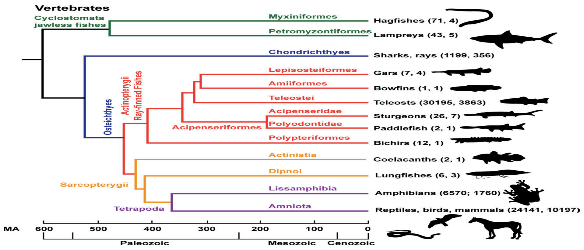


Astronomy has become one of the first digital science, replacing photographs with digital images.

The Large Synoptic Survey Telescope (LSST) has a 3.2 gigapixel camera and will produce 15-20 terabytes of data per night and more than ***100 petabytes*** *over its first 10 years of operation.*

With the genomic revolution, biology and biomedicine are rapidly becoming digital sciences. The opportunities for breakthroughs in these areas are just beginning to be explored as exemplified by the Genome 10K project.

GENOME 10K.

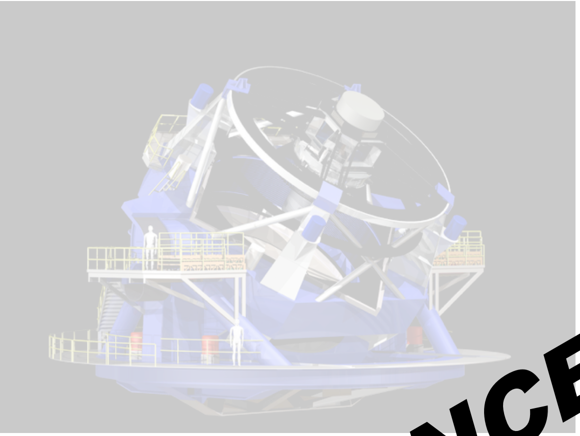


Petaflops & Petabytes

Who Needs Petabytes?

SCIENCE, 3 MARCH 2017

Researchers propose to sequence 1,000,000 eukaryote genomes

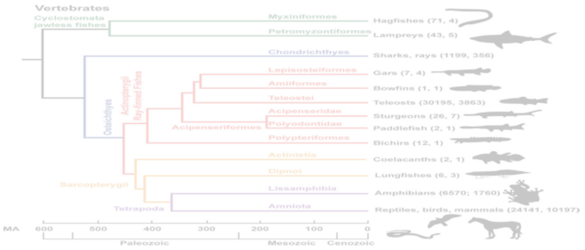


Astronomy has become one of the first digital science, replacing photographs with digital images.

*The Large Synoptic Survey Telescope (LSST) has a 3.2 gigapixel camera and will produce 15-20 terabytes of data per night and more than **100 petabytes** over its first 10 years of operation.*

With the genomic revolution, biology and biomedicine are rapidly becoming digital sciences. The opportunities for breakthroughs in these areas are just beginning to be explored as exemplified by the Genome 10K project.

GENOME 10K.



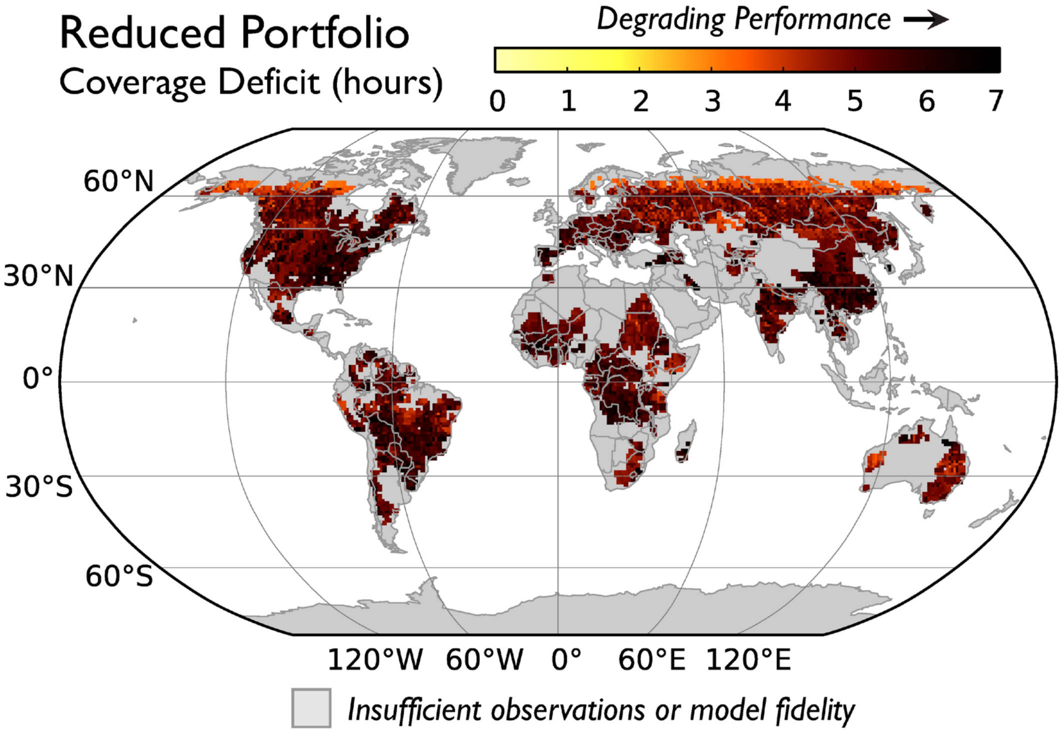
Petaflops & Petabytes

Enabling Entirely New Applications

Optimization of Satellite Constellations

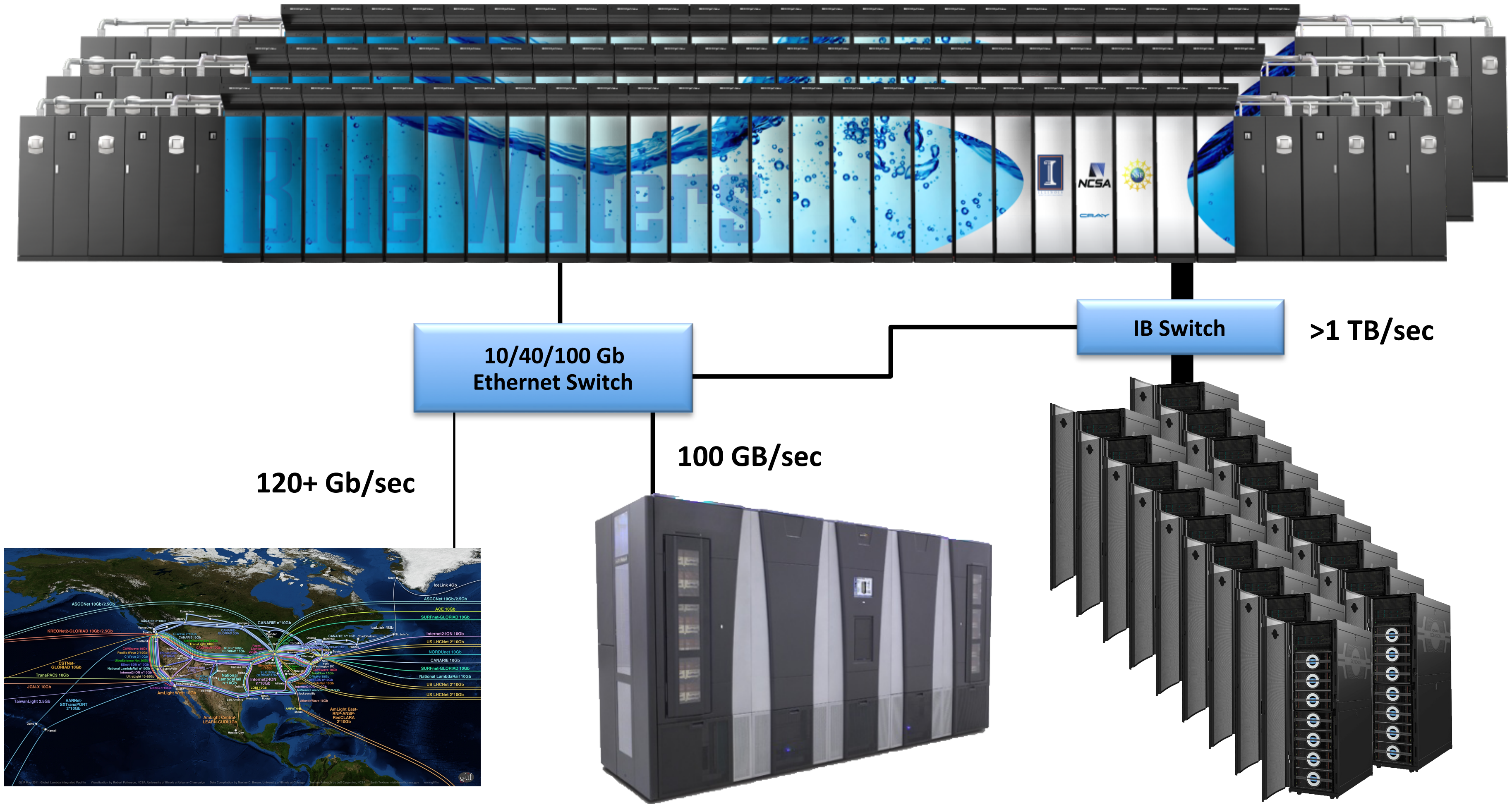
This project will enable the scientific and space agency communities to optimize the architectures of future satellite constellations to ensure that they deliver high-fidelity data for a broad array of environmental research applications.

P. Reed (Cornell), E. F. Wood (Princeton), M. Ferringer (Aerospace Corp.)



Petaflops & Petabytes

Blue Waters Petascale Computing System



10/40/100 Gb Ethernet Switch

IB Switch

>1 TB/sec

120+ Gb/sec

100 GB/sec

WAN

Spectra Logic: 300 PBs

Sonexion: 26 PBs

Petaflops & Petabytes

Specifications: Blue Waters & Titan

	Blue Waters	Titan
Vendor(s)	Cray/AMD/NVIDIA	
Processors	Interlagos/Kepler	Interlagos/Kepler
Peak Performance	**13.1 PF**	**27.1 PF**
CPU/GPU	7.6 / 5.5 PF	2.6 / 24.5 PF
Number of Chips (CPU/GPU)	48,352/4,224	18,688/18,688
Amount of Memory	1.66 PB	0.71 PB
Disk Storage, Capacity (usable) Disk Storage, Bandwidth (sustained)	26 PB 1.2 TB/s	>10 TB 0.24 TB/s
Archival Storage, Capacity (usable) Archival Storage, Bandwidth (sustained)	300 PB ~100 GB/s	125 PB 18 GB/s

Petaflops & Petabytes

Specifications: Blue Waters & Titan

	Blue Waters	Titan
Vendor(s)	Cray/AMD/NVIDIA	
Processors	Interlagos/Kepler	Interlagos/Kepler
Peak Performance	13.1 PF	27.1 PF
CPU/GPU	7.6 / 5.5	2.6 / 24.5
Number of Chips (CPU/GPU)	48,352/4,224	18,688/18,688
Amount of Memory	1.66 PB	0.71 PB
Disk Storage, Capacity (usable)	26 PB	>10 TB
Disk Storage, Bandwidth (sustained)	1.2 TB/s	0.24 TB/s
Archival Storage, Capacity (usable)	300 PB	125 PB
Archival Storage, Bandwidth (sustained)	100 GB/s	18 GB/s

CAUTIONARY NOTE #1:

Sustained Performance on Blue Waters Benchmark Suite

Blue Waters: 1.37 PFs

Titan: 0.64 PFs

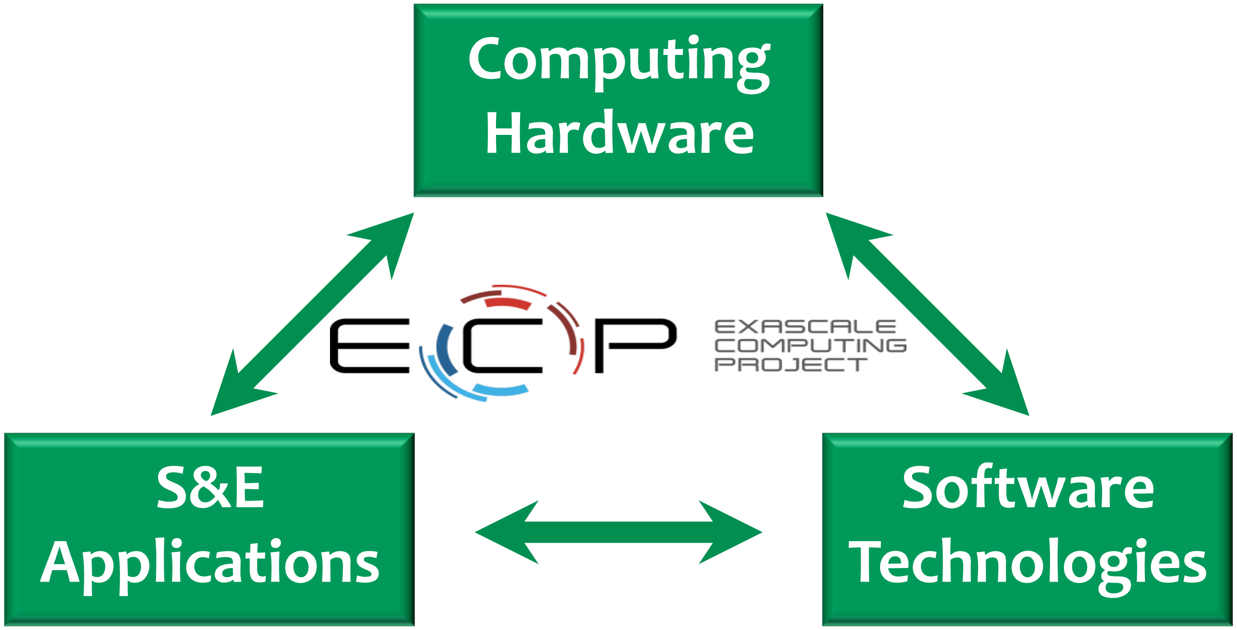
Real-world performance depends on applications taking advantage of hardware innovations

Moving to the Exascale: Exascale Computing Project

The U.S. Department of Energy has embarked on an ambitious program to develop, with industry, an exascale computer that can support a broad range of science & engineering applications. The project combines hardware innovations with the development of critical software technologies and science & engineering applications.

Moving to the Exascale

Integrated Approach to Advancing Computing



Chemistry • Climate/Geophysics
Accelerators • Biosciences • Subsurface
Astrophysics/Cosmology • Fusion energy
Energy systems • Energy devices
… • High energy physics

Node OS • Runtimes • Systems software
Programming models • Math libraries
Visualization • Data analysis • IO
Communications Libraries • Workflow
Resilence • …

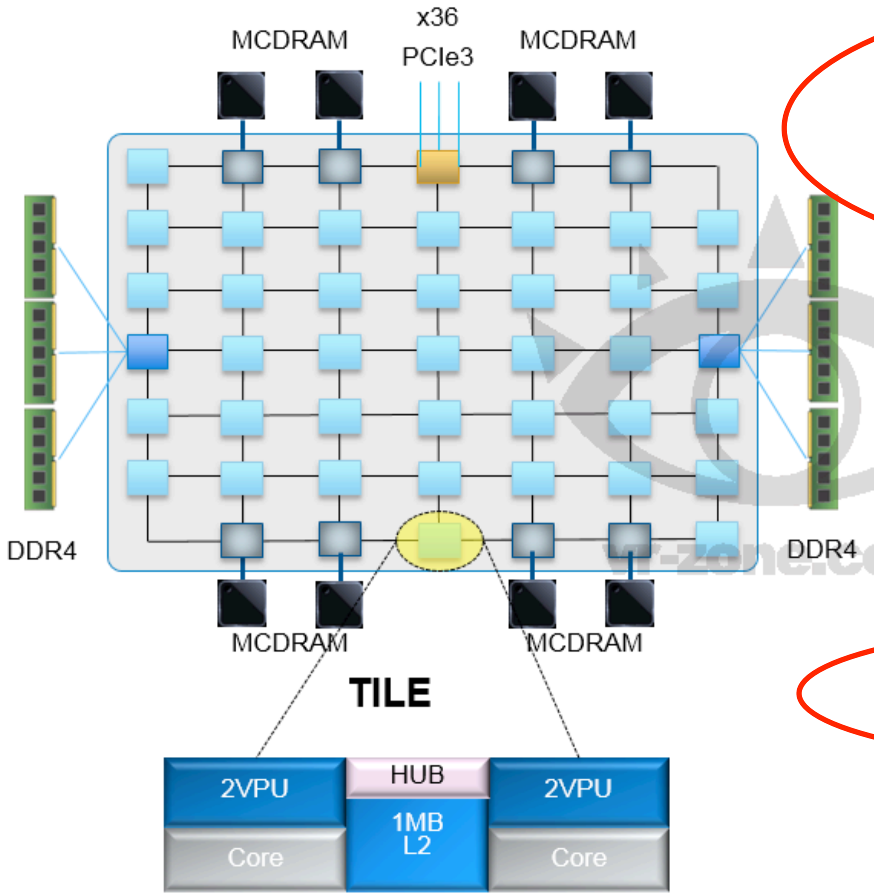
Moving to the Exascale

Oak Ridge's Summit & Argonne's Aurora Systems

	Summit (2018)	Aurora (2018)
Processor	IBM Power9/NVIDIA Volta	Intel Knights Hill
Peak Performance	>150 PF	180 PF
Cores/Processor	Up to 24	>72
Number of Nodes	~3,400	>50,000
Memory	>1.7 PB	>7 PB
Interconnect BS Bandwidth	?	>500 TB/s
File System Capacity	~120 PB	>150 PB
File System Bandwidth	~1 TB/s	>1 TB/s
Peak Power	~ 10 MW	13 MW

Moving to the Exascale

Knights Landing Architecture



Up to 72 Intel Architecture cores based on Silvermont (Intel® Atom processor)

- Four threads/core
- Two 512b vector units/core
- Up to 3x single thread performance improvement over KNC generation

Full Intel® Xeon compatibility thrc TSX)

6 channels of DD 384GB

36 lanes PCI Express* Gen 3

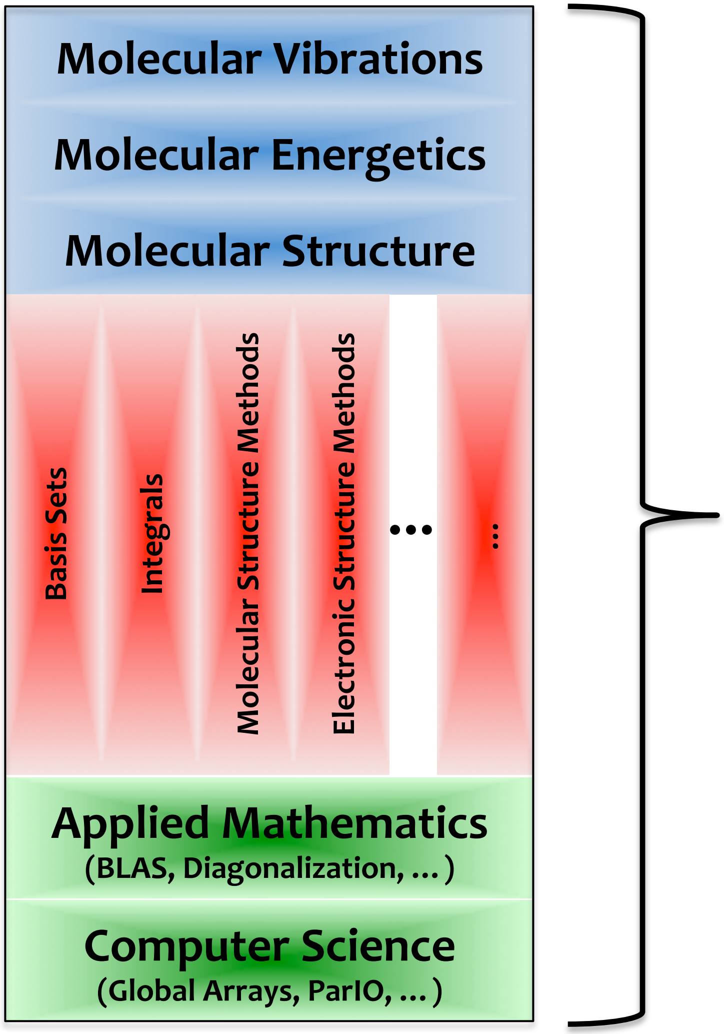
8/16GB of high-bandwidth on-package MCDRAM memory >500GB/sec

200W TDP

New opportunities,
New challenges

NWChemEx Project

NWChem: An Exemplary SC Application



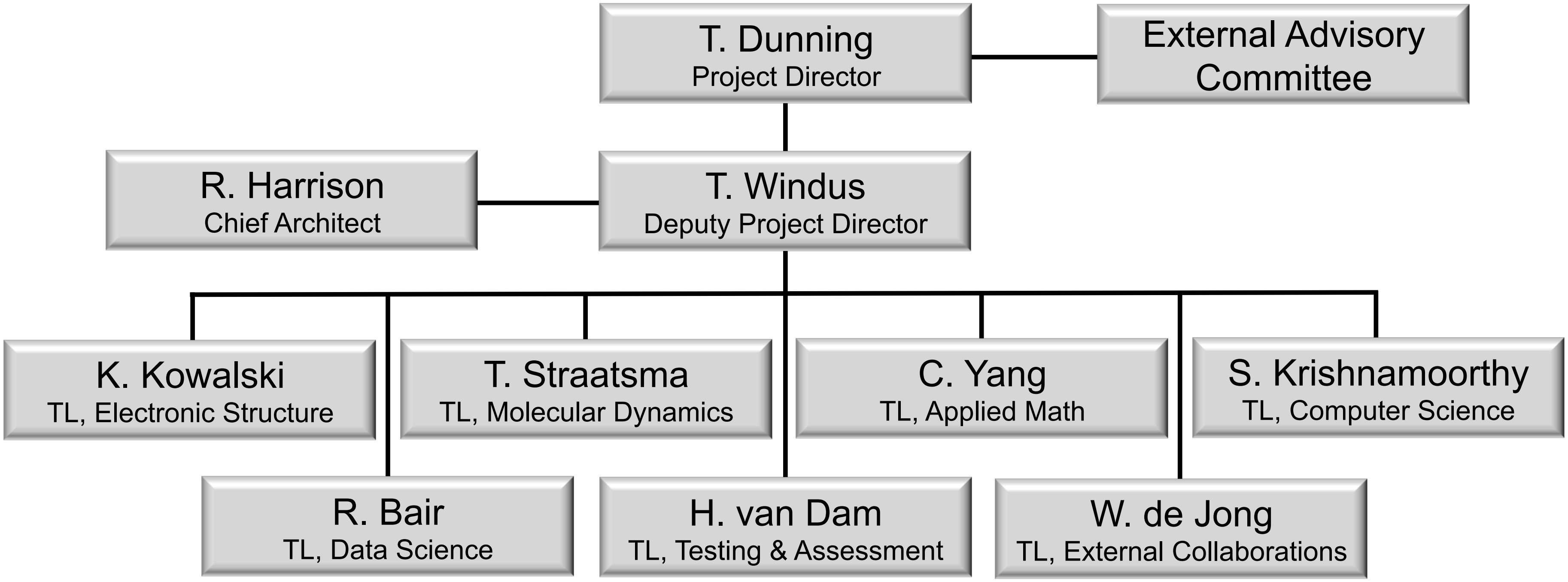
- **NWChem Team**
  - Computational chemists
  - Computer scientists
  - Applied mathematicians
- **Current Status**
  - Implements broad range of electronic structure and molecular dynamics methods
  - Approx. 4 million lines of code (3 million generated by TCE)
  - Written in Fortran, beginning in 1990s

NWChemEx Project

Goals of NWChemEx Project

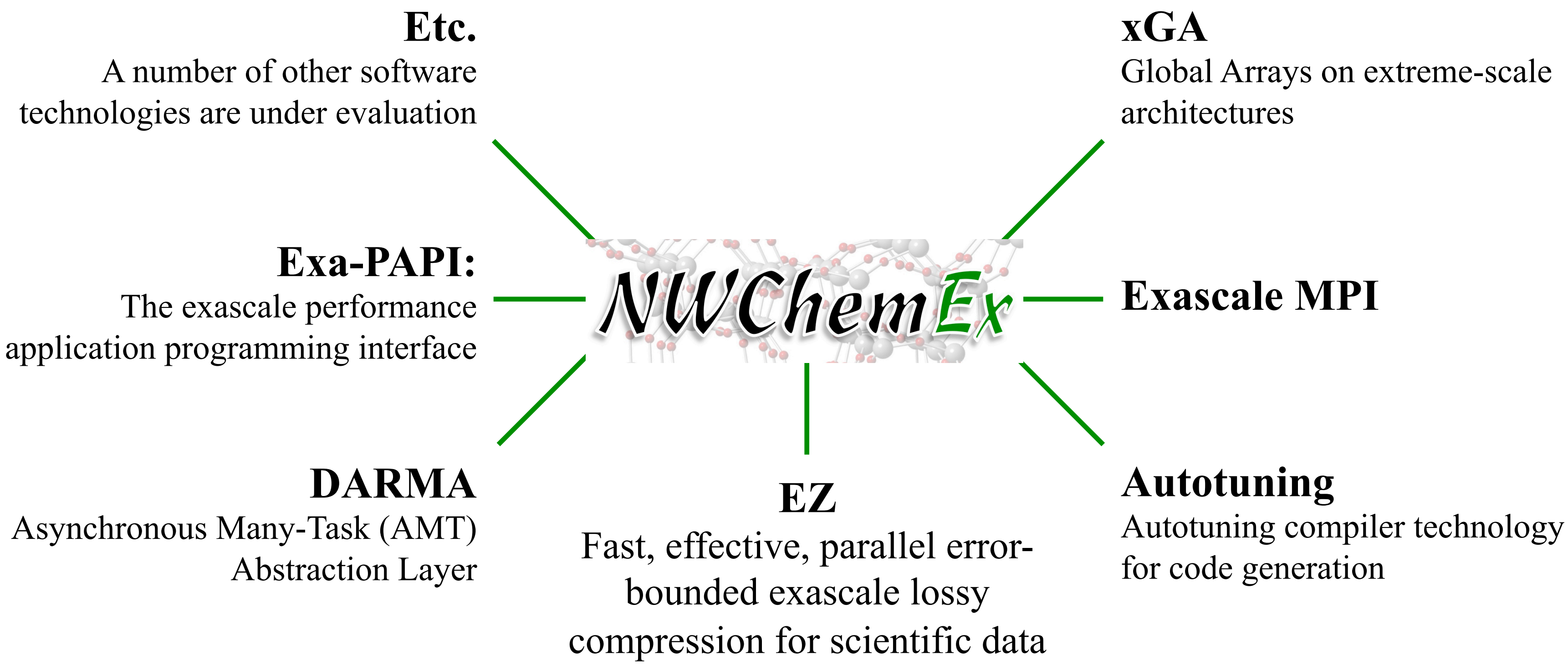
- **Redesign and re-implement (in C++) NWChem** for exascale computing technologies
- **Provide molecular modeling capabilities** needed to address two decadal science challenges:
  - Design of feedstock for the efficient production of biomass
  - Design of new catalysts for the efficient conversion of biomass-derived intermediates into biofuels
- **Provide framework for community effort** to develop next-generation molecular modeling package that supports broad range of chemistry research on computing systems ranging from terascale workstations and petascale servers to exascale computers

Organization of NWChemEx Project



- Six national laboratories, one university
- Eighteen staff (none full time) plus support staff
- Postdoctoral fellows (TBD)

Integration of NWChemEx and ECP Projects



Etc.
A number of other software technologies are under evaluation

Exa-PAPI:
The exascale performance application programming interface

DARMA
Asynchronous Many-Task (AMT) Abstraction Layer

EZ
Fast, effective, parallel error-bounded exascale lossy compression for scientific data

xGA
Global Arrays on extreme-scale architectures

Exascale MPI

Autotuning
Autotuning compiler technology for code generation

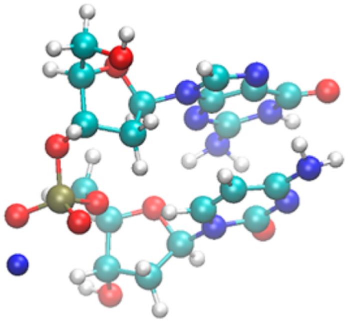
NWChemEx Project

Measuring Performance of NWChem #1

Method	Time(s)*	GFLOP Count	PF/s
(T)	5024	5,948,249,197	1.18

* On 20,000 XE6 nodes (Blue Waters)

V. M. Anisimov, G. H. Bauer, K. Chadalavada, R. M. Olson, J. Glenski, W. T. C. Kramer, E. Aprà, and K. Kowalski, *J. Chem. Theory Comput.* **10**, 4307-4316 (2014).



guanine− cytosine deoxydinucleotide monophosphate + Na^+

- NWChem achieves impressive performance on petascale computers for the most flop-intensive calculations
- For CCSD(T) calculations, which is the current “gold” standard, this is the (T) algorithm
- **NWChem achieves over 1 PF/s on 20,000 nodes of Blue Waters on (T) algorithm**

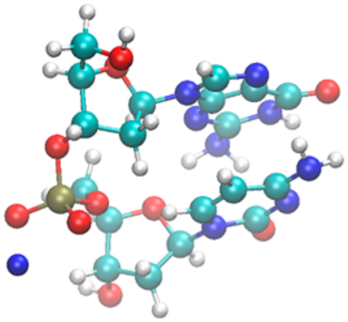
NWChemEx Project

Measuring Performance of NWChem #2

Method	Time(s)*	GFLOP Count	PF/s
CCSD	29,500	195,796,351	0.005
(T)	5024	5,948,249,197	1.18
CCSD(T)	34,524	6,144,045,548	0.18

* On 20,000 XE6 nodes (Blue Waters)

V. M. Anisimov, G. H. Bauer, K. Chadalavada, R. M. Olson, J. Glenski, W. T. C. Kramer, E. Aprà, and K. Kowalski, *J. Chem. Theory Comput.* **10**, 4307-4316 (2014).



guanine− cytosine deoxydinucleotide monophosphate + Na^+

- Need CCSD amplitudes for the (T) algorithm
- CCSD algorithm is far more complex with a much higher communication/compute ratio than the (T) algorithm
- **CCSD algorithm consumes 85% of the time, lowering the overall performance to just 0.18 PF/s**

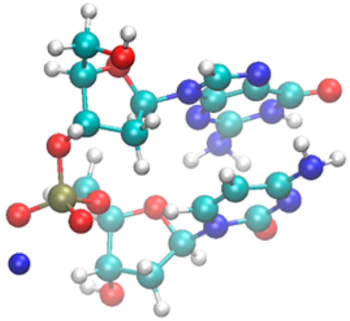
NWChemEx Project

Measuring Performance of NWChem #3

Method	Time(s)*	GFLOP Count	PF/s
CCSD	14,406	195,796,351	0.01
(T)	5024	5,948,249,197	1.18
CCSD(T)	19,430	6,144,045,548	0.32

* On 20,000 XE6 nodes (Blue Waters)

V. M. Anisimov, G. H. Bauer, K. Chadalavada, R. M. Olson, J. Glenski, W. T. C. Kramer, E. Aprà, and K. Kowalski, *J. Chem. Theory Comput.* **10**, 4307-4316 (2014).



guanine− cytosine deoxydinucleotide monophosphate + Na^+

- Analysis of communications patterns in CCSD algorithm found that access to ST2 array, stored in global memory, is responsible for performance bottleneck
- Replicating ST2 array dramatically reduced communications wait time
- **Tradeoff: ST2 is so large, only 1 core of 16 could be used, although new algorithm is still nearly 2x faster**

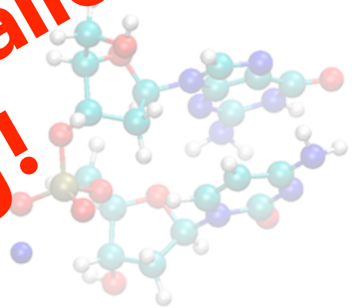
NWChemEx Project

Performance of NWChem on Blue Waters II

Method	Time(s)*	GFLOP Count	PF/s
CCSD	14,406	195,796,[illegible]	0.01
(T)	5024	5,9[illegible]8,249,197	1.18
CCSD(T)	19,430	6,144,045,548	0.32

* On 20,000 XE6 nodes (Blue Waters)

V. M. Anisimov, G. H. Bauer, K. Chadalavada, R. M. Olson, J. Glenski, W. T. C. Kramer, E. Aprà, and K. Kowalski, *J. Chem. Theory Comput.* 10, 4307-4316 (2014).



guanine− cytosine deoxydinucleotide monophosphate + Na^+

CAUTIONARY NOTE #2:

This example illustrates the challenge of exascale computing!

So, the CCSD algorithm consumes ¾-th of the time. Further, the algorithm uses a substantial amount of memory, duplicating arrays to minimize communication costs, which limits the number of cores/node that can be used—just 1 of 16 cores on a Blue Waters node that has 64 GBs of memory on the node.

Substantial advances in performance are only possible if new algorithms or approaches can overcome system limitations

Déjà vu: SciDAC 2000

Peak Performance Skyrocketing

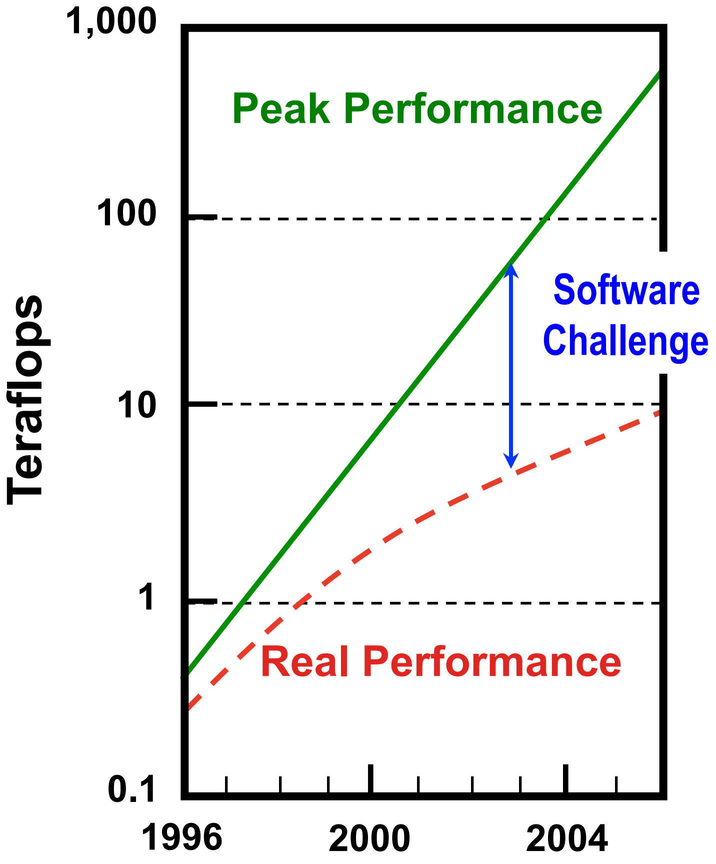
- In past 10 years, peak performance has increased 100x; in next 5 years, it will increase at least 100x

Real Performance Increasing, But ...

- Efficiency has declined from 30-40% on vector supercomputers of 1990s to as little as 5-10% on parallel supercomputers of today

Research Challenge: Software

- Scientific codes to model and simulate physical processes and systems
- Computing and mathematics software to enable use of advanced computers for scientific applications
- Continuing problem as computer architectures undergo fundamental changes



In Summary: What Do you Want?

Just a phone?

Or a smart phone?

In Summary: What Do you Want?

Just a phone?

Or a smart phone?

For Supercomputers:

Applications are not just value-added, they are the value

Thank You!