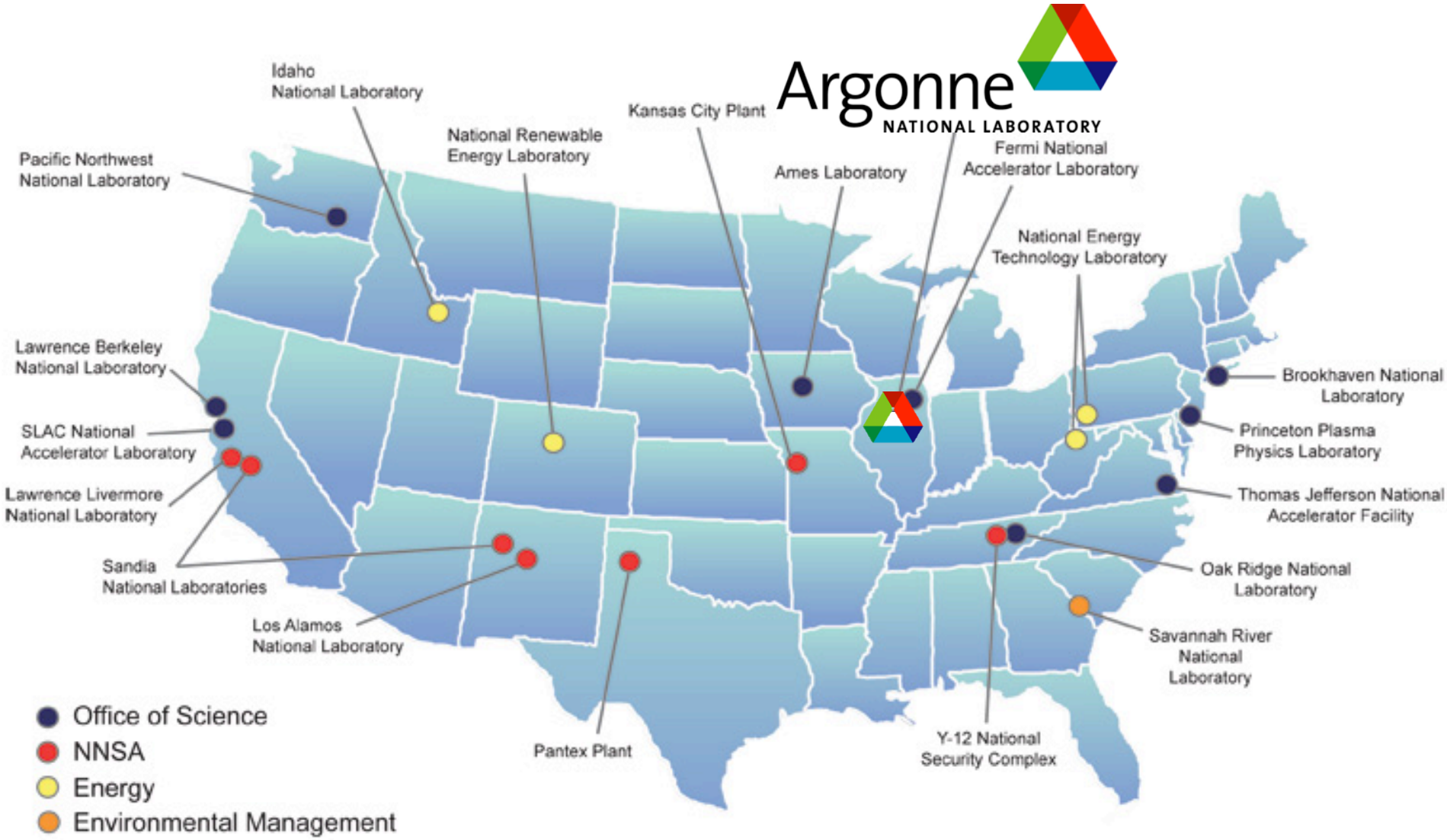
Toward Extreme Scale

Marc Snir
Director, Mathematics and Computer Science Division
Argonne National Laboratory

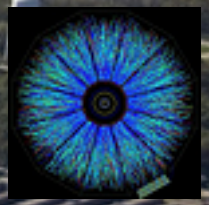
Professor, Dept. of Computer Science, UIUC

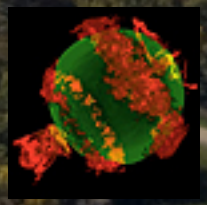
Argonne: Vital part of DOE National Laboratory System

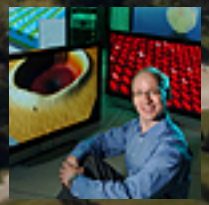


About Argonne

- $675M operating budget
- 3,200 employees
- 1,450 scientists and engineers
- 750 Ph.D.s







Direct descendent of Enrico Fermi's Metallurgical Laboratory

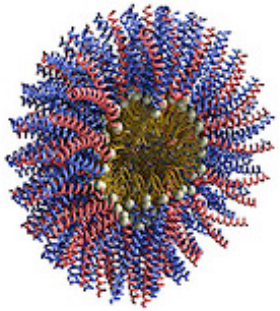
Argonne's mission: To provide science-based solutions to pressing global challenges

Through discovery and transformational science and engineering…

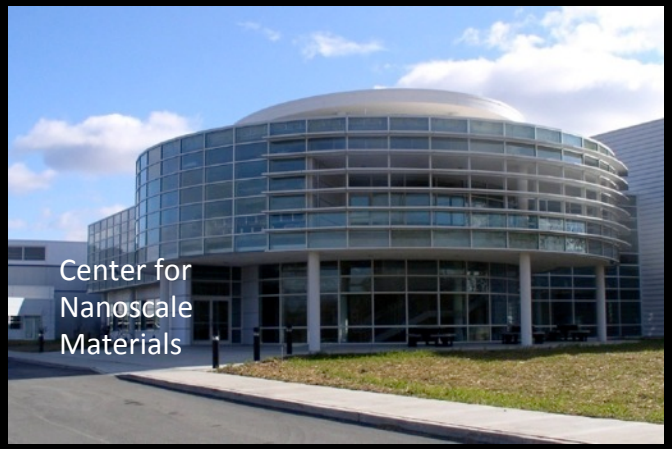
- World-leading hard x-ray sciences & sources
- Discovery science for energy
- Leadership computing and computational ecosystem
- Fundamental physics and accelerator capabilities
- Materials & systems engineering solutions

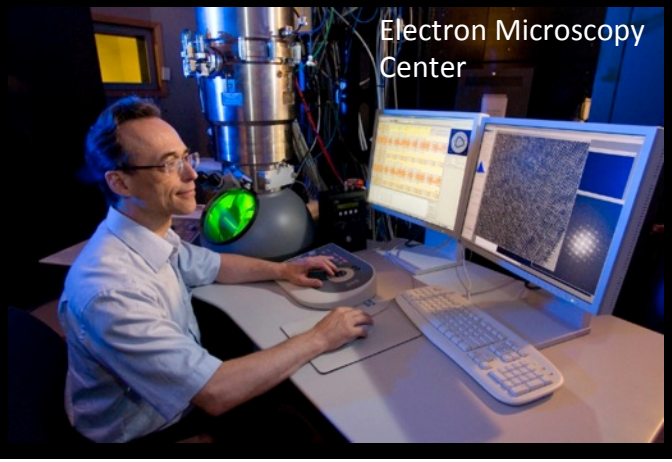
and through use-inspired science and engineering

- Energy Storage
- Sustainable Transportation
- Nuclear Energy
- Environmental Genomics
- National Security



Major Scientific User Facilities





Improving Beijing air quality for Olympic Games

Available online at www.sciencedirect.com
ScienceDirect
Atmospheric Environment 41 (2007) 480–492
ATMOSPHERIC ENVIRONMENT
www.elsevier.com/locate/atmosenv
ELSEVIER

Air quality during the 2008 Beijing Olympic Games

David G. Streets[a,*], Joshua S. Fu[b], Carey J. Jang[c], Jiming Hao[d], Kebin He[d], Xiaoyan Tang[e], Yuanhang Zhang[e], Zifa Wang[f], Zuopan Li[b], Qiang Zhang[a,d], Litao Wang[d], Binyu Wang[c], Carolyne Yu[a]

[a]Decision and Information Sciences Division, Argonne National Laboratory, Argonne, IL 60439, USA
[b]Department of Civil & Environmental Engineering, University of Tennessee, Knoxville, TN 37996, USA
[c]Office of Air Quality Planning & Standards, US Environmental Protection Agency, Research Triangle Park, NC 2
[d]Department of Environmental Science and Engineering, Tsinghua University, Beijing 100084, PR China
[e]College of Environmental Sciences, Peking University, Beijing 100871, PR China
[f]Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, PR China

Received 13 April 2006; received in revised form 14 July 2006; accepted 19 August 2006

Abstract

China is taking major steps to improve Beijing's air quality for the 2008 Olympic Games. How
fine particulate matter and ozone in Beijing often exceed healthful levels in the summertime.
Models-3/CMAQ model simulation over the Beijing region, we estimate that about 34% of PM_2
of ozone during high ozone episodes at the Olympic Stadium site can be attributed to sources ou
Hebei and Shandong Provinces and the Tianjin Municipality all exert significant influence on
sustained wind flow from the south, Hebei Province can contribute 50–70% of Beijing's $PM_{2.5}$
of ozone. Controlling only local sources in Beijing will not be sufficient to attain the air qu
Olympics. There is an urgent need for regional air quality management studies and new

- Tsinghua University
- Peking University
- Chinese Academy of Sciences
- Argonne National Laboratory
- U.S. Environmental Protection Agency
- University of Tennessee

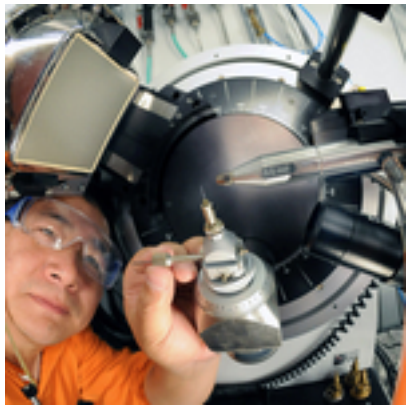
Beijing scene after rain storm

Argonne National Laboratory: 4 Directorates

Computing, Environment & Life Sciences (CELS)

Energy Engineering & Systems Analysis (EESA)

Photon Sciences (PS)



Physical Sciences & Engineering (PSE)

CELS: 4 Divisions

The Argonne Leadership Computing Facility

Mira: 10 PF BG/Q, ~750,000 cores, 0.75 PB

Biosciences Division

Environmental Science Division

Mathematics and Computer Science Division

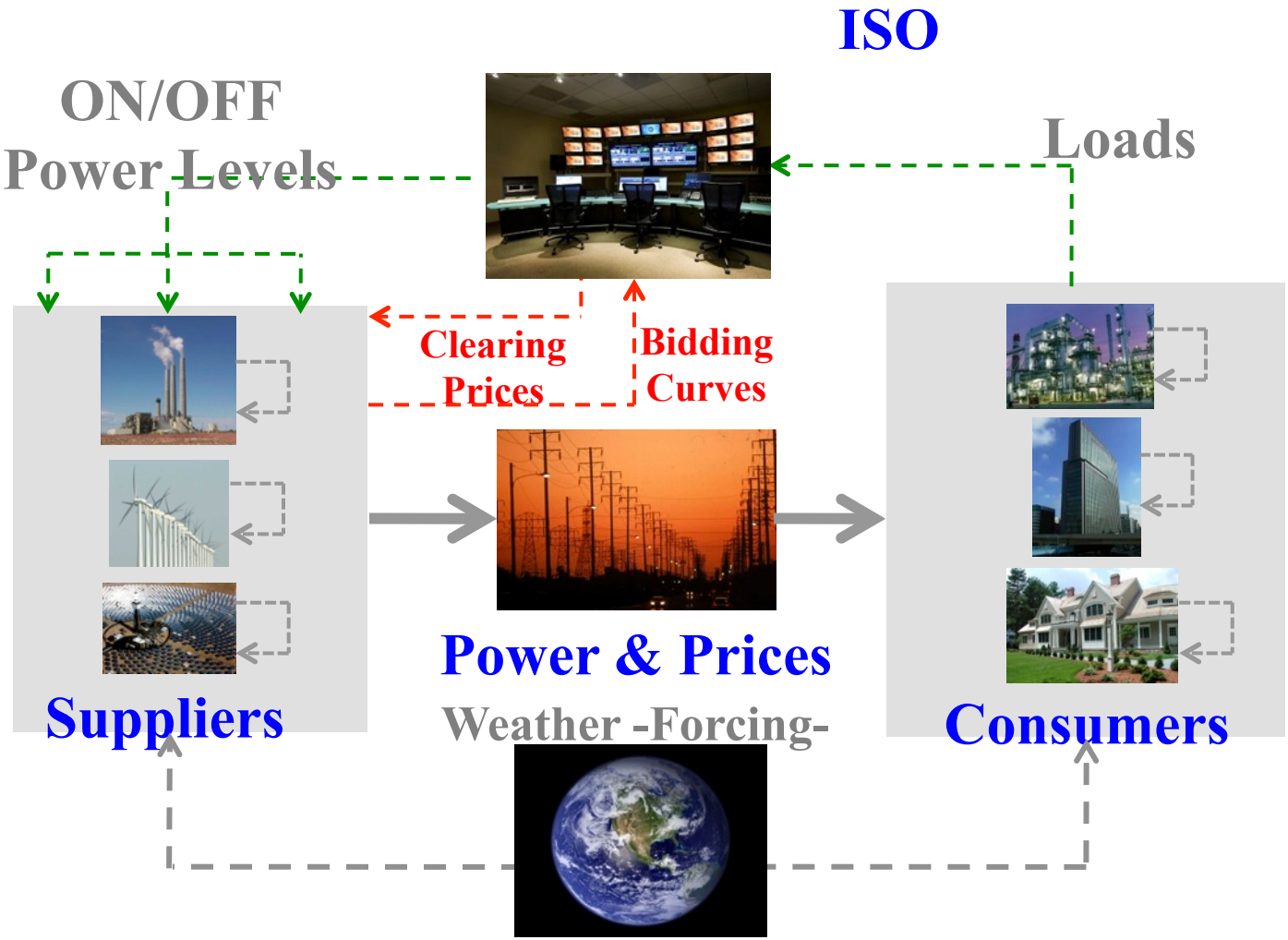
MCS

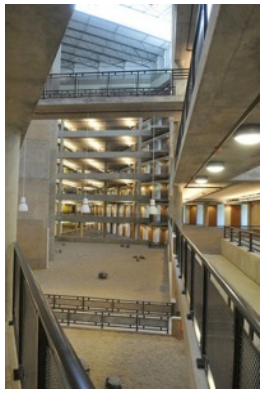
- **Extreme Computing Group**: Software infrastructure for extreme scale computers
- **Big Data Group**: Software infrastructure for storage, communication and analysis of large & complex data sets
- **Applied Math Group**: Scalable numerical algorithms and scientific libraries
  - Sparse linear algebra (PETSc), PDE solvers (MOAB), Optimization (TAO)
- **Computational Science Group**: Application of advanced compute methods to selected application areas
  - Bioinformatics, climate modeling, nuclear engineering, cosmology
- **Computational Institute** (joint with U. of Chicago): Collaborative environments
  - Grid, cloud
- ~100 people
- Collaborations with China, Germany, France…

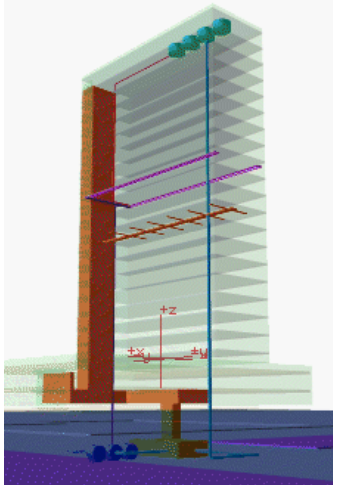
Math

A Few Examples

Stochastic Optimization (Zavala)





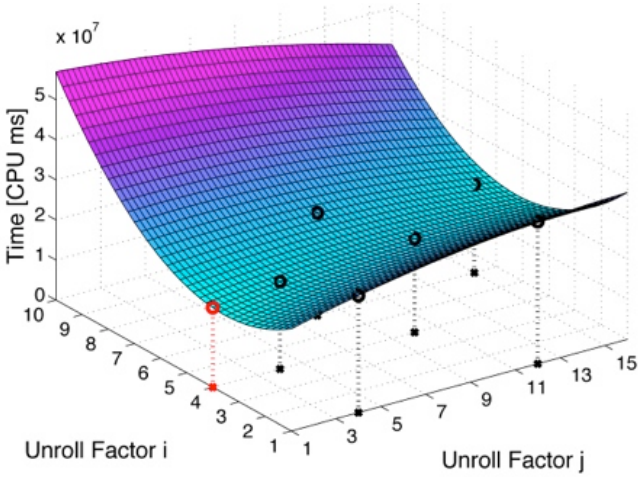


Decide when to turn on & off power stations

Control temperature in a (smart) building

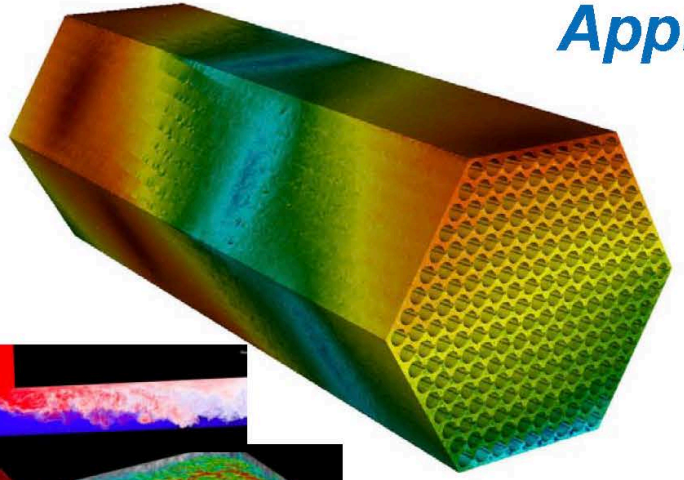
Optimization (Wild)

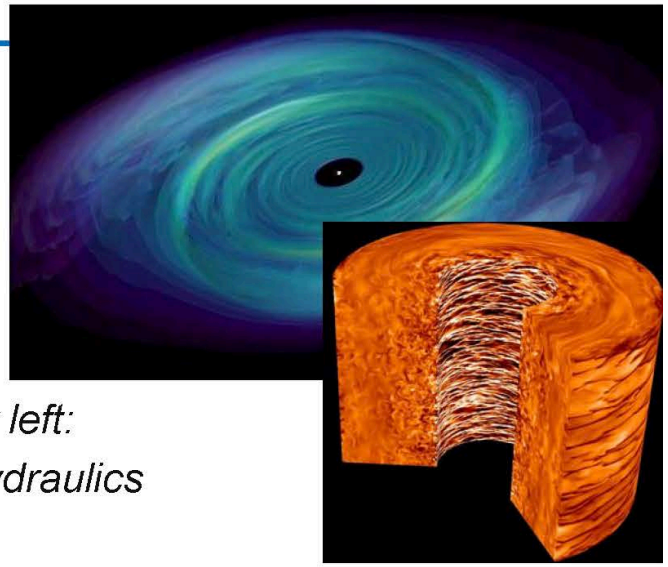
- Autotuning
  - New algorithms tailored for auto-tuning problems: mixed integer and continuous parameters
  - Noisy function evaluations, disjoint local minima
  - Incorporate existing tuning heuristics in global search strategies

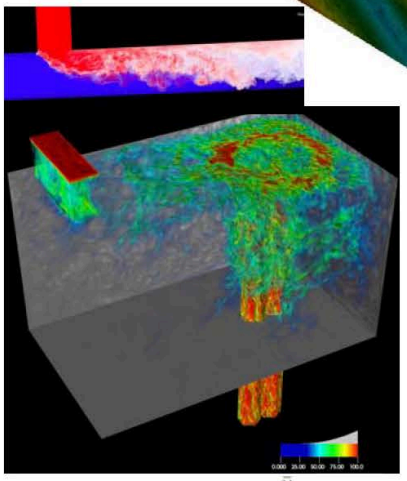


CFD – Nek5000

Applications

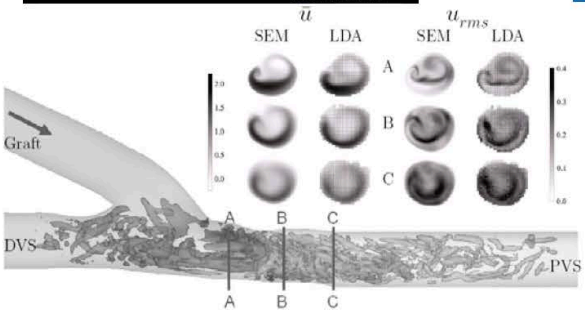


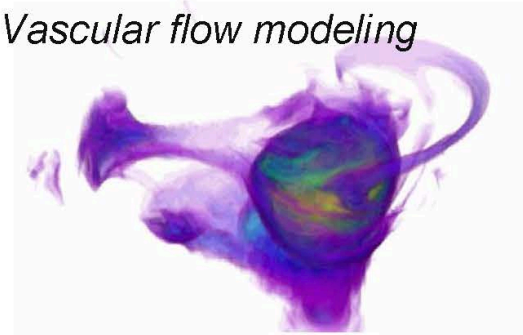


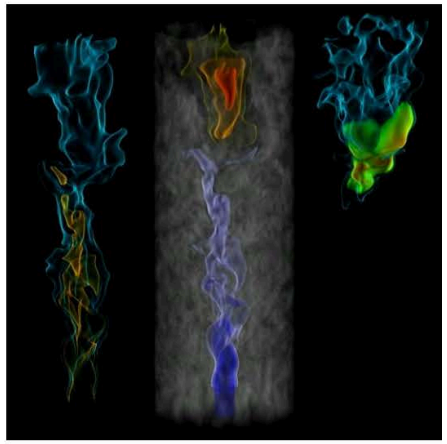


Clockwise from upper left:

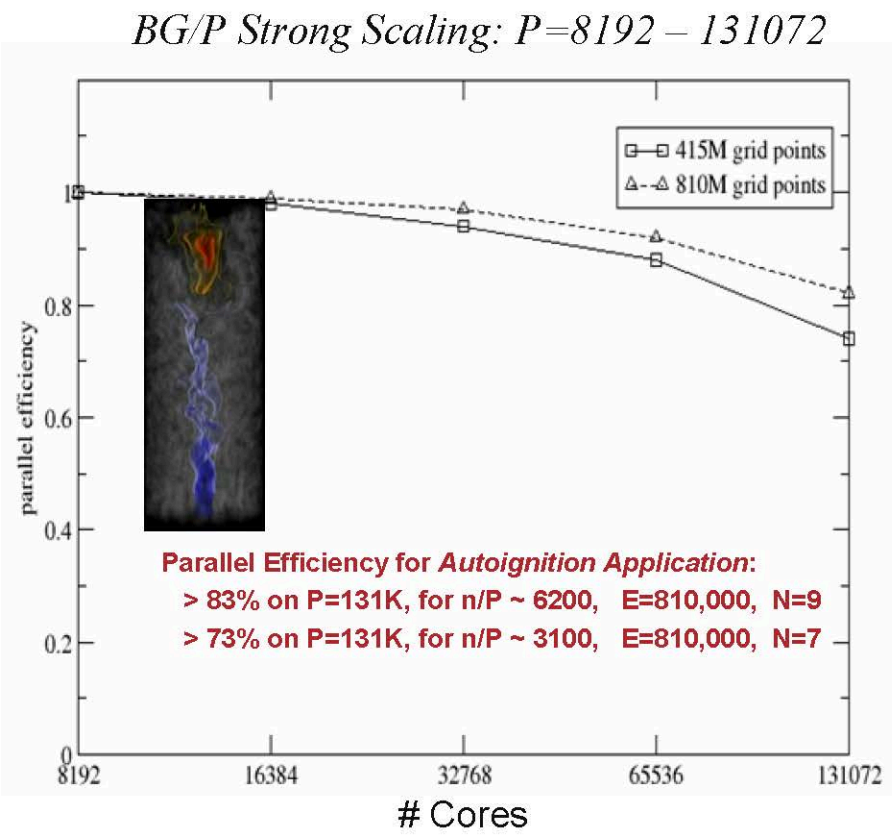
- *Reactor thermal-hydraulics*
- *Astrophysics*
- *Combustion*
- *Oceanography*
- *Vascular flow modeling*

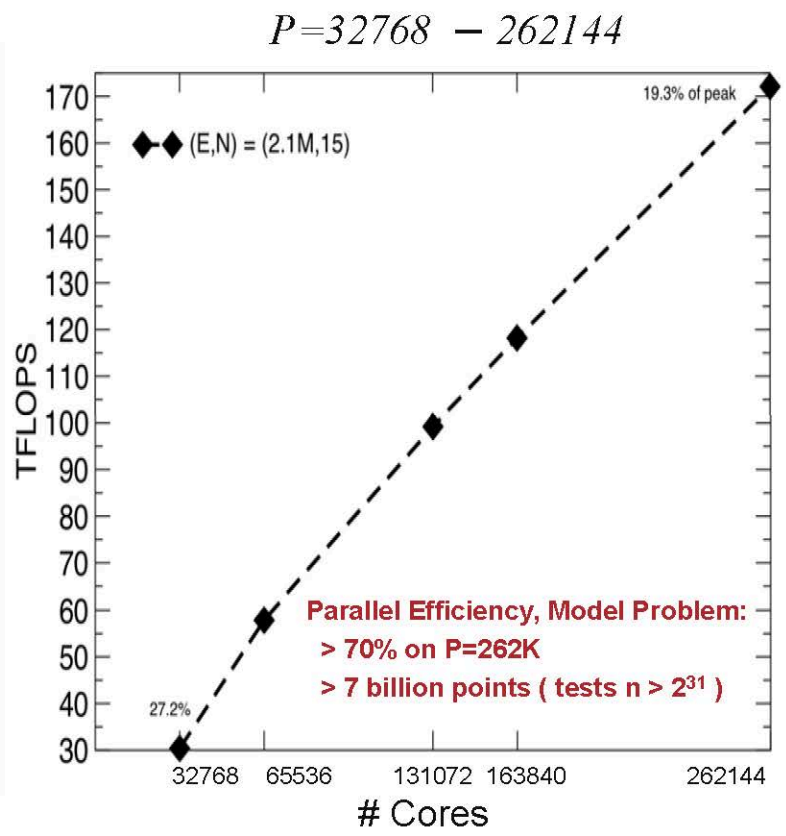






Nek5000 scalability

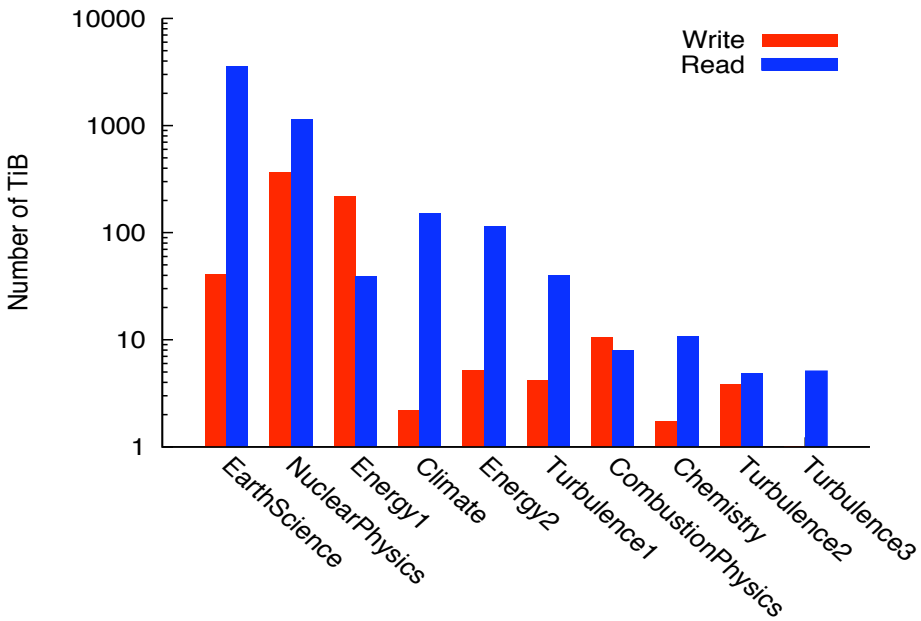




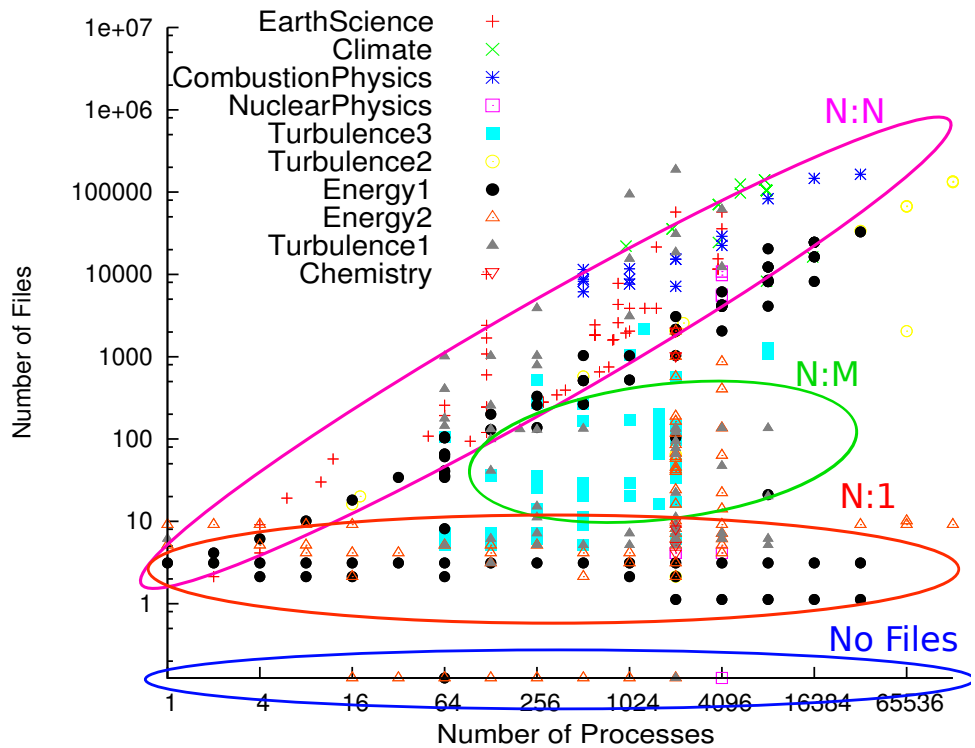
Big Data

Examples

Understanding I/O patterns (Darshan)

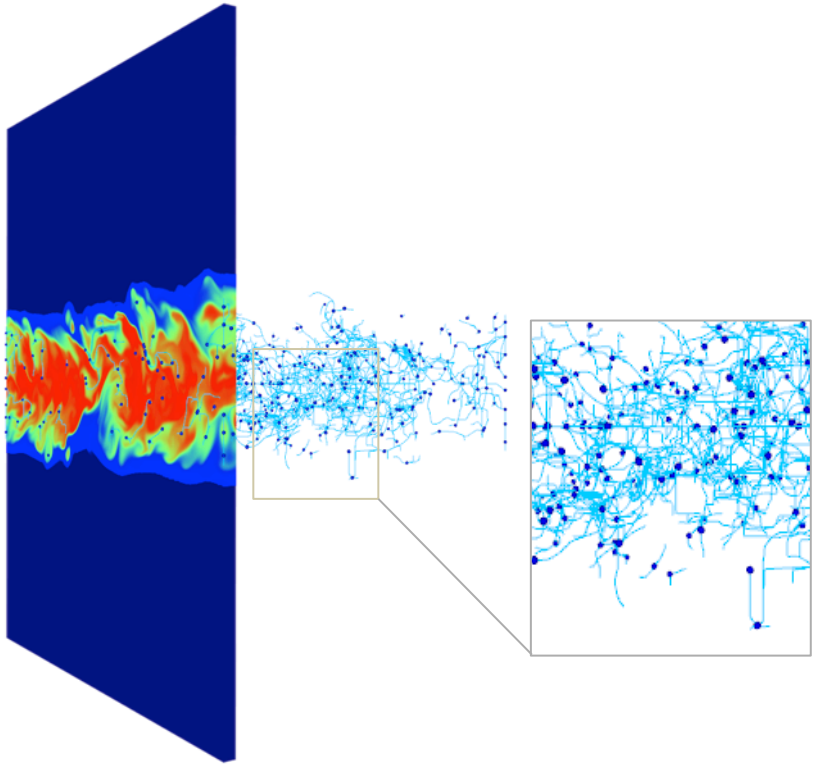


Reads vs. Writes

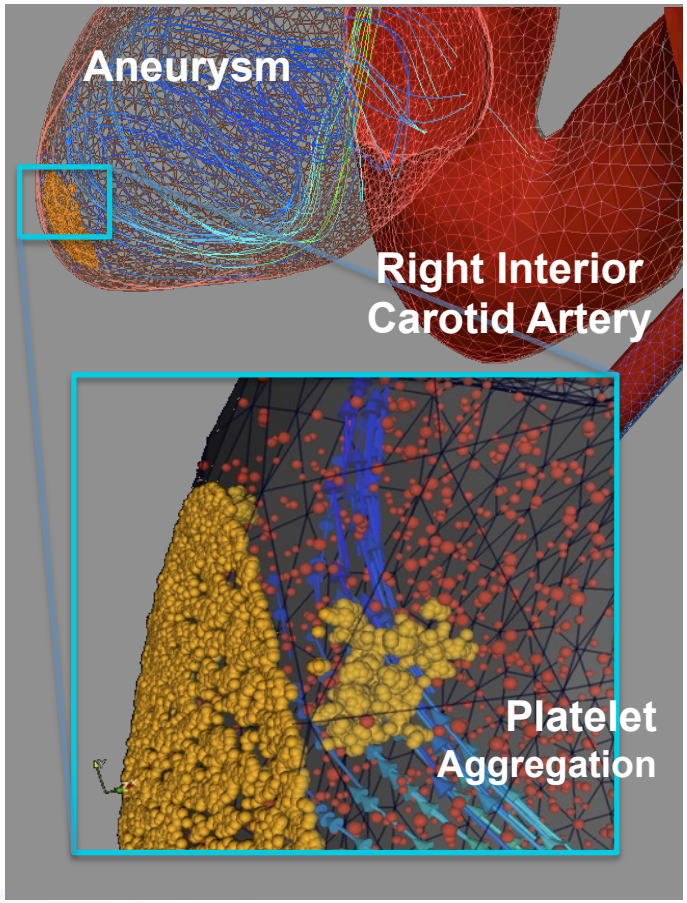


files

Analyzing and Displaying Complex data



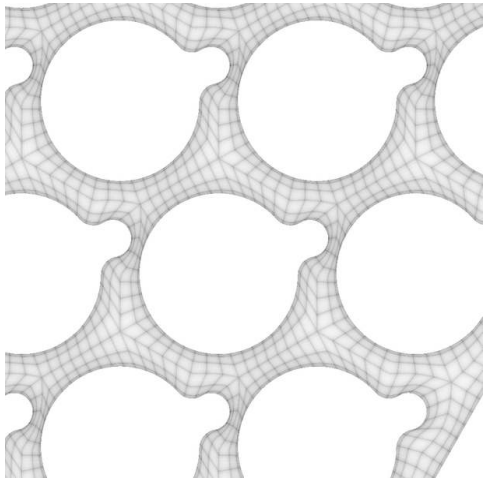
- Atomistic data representations for plasma, red blood cells, and platelets from MD simulation.
- Field data for ensemble average hydrodynamics solution



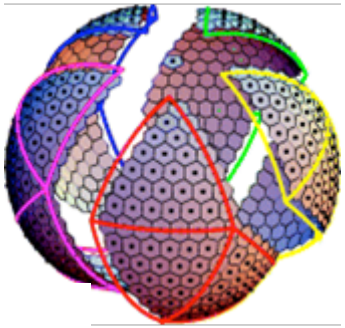
- Morse-Smale complex graph created from gradient field (combustion)

Developing Future Storage Systems that Support Computational Science Data Models

Motif	Data Model/ Structure
Dense Linear Algebra	Multidimensional Arrays
Sparse Linear Algebra	Sparse Matrix
Spectral Methods	Multidimensional Arrays
N-Body Methods	Trees, Unstructured Meshes
Structured Grids (+ AMR)	Multidimensional Arrays
Unstructured Grids (+AMR)	Unstructured Meshes
Graph Traversal	Sparse Matrix, DAG



Cross-section of spectral element mesh used in large eddy simulation of 217-pin reactor subassembly.



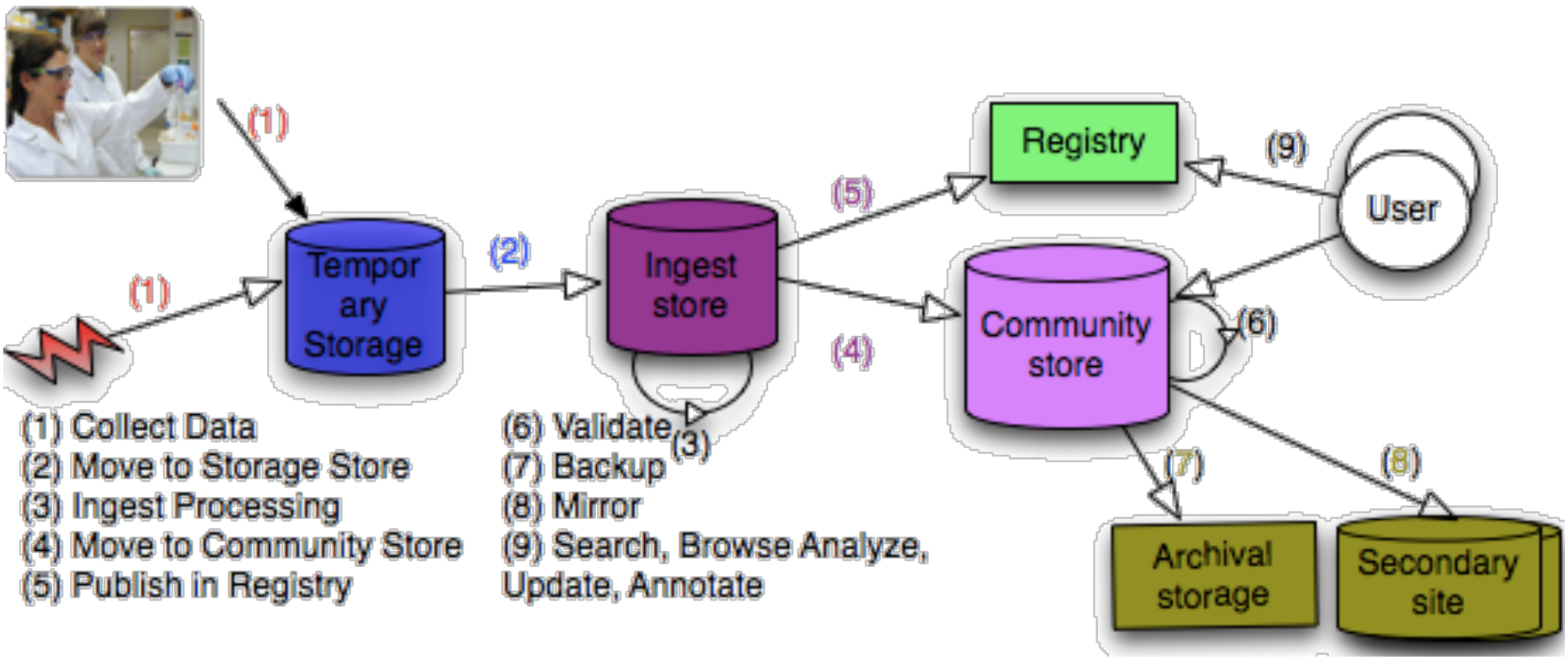
Geodesic grid used in global climate resolving model.

Collaborative Environments

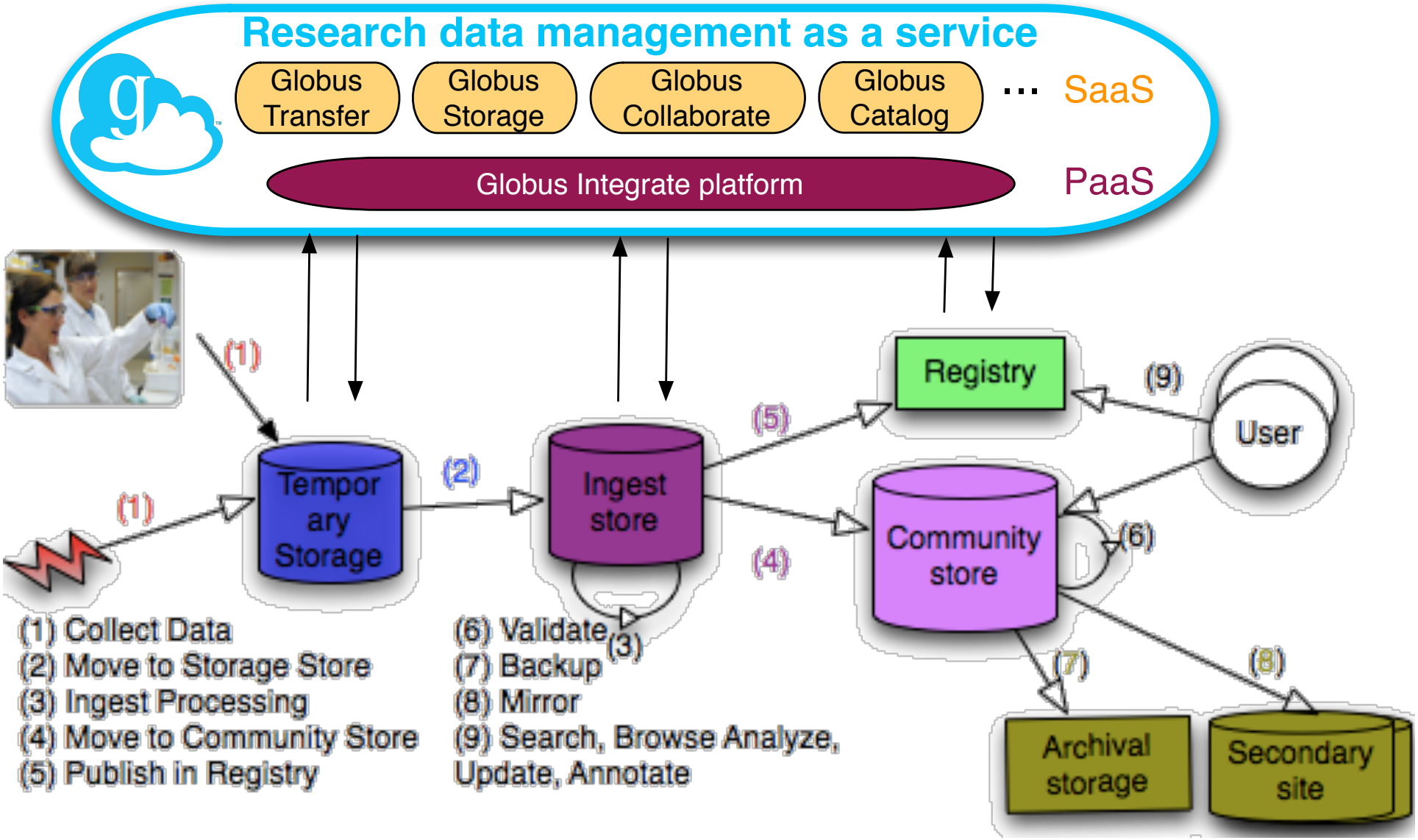
Examples

Towards "research IT as a service"

- Dark Energy Survey
- Galaxy genomics
- LIGO observatory
- SBGrid structural biology consortium
- NCAR climate data applications
- Land use change; economics

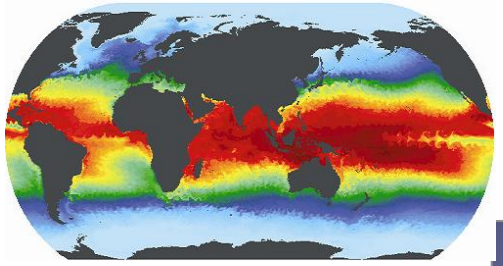


Towards "research IT as a service"

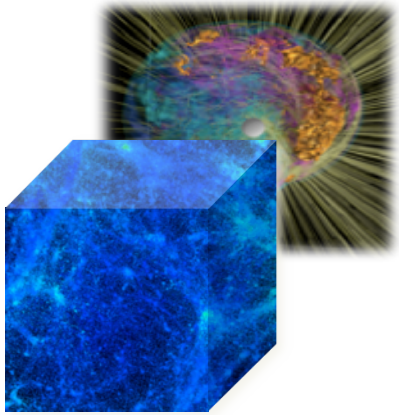


Exascale

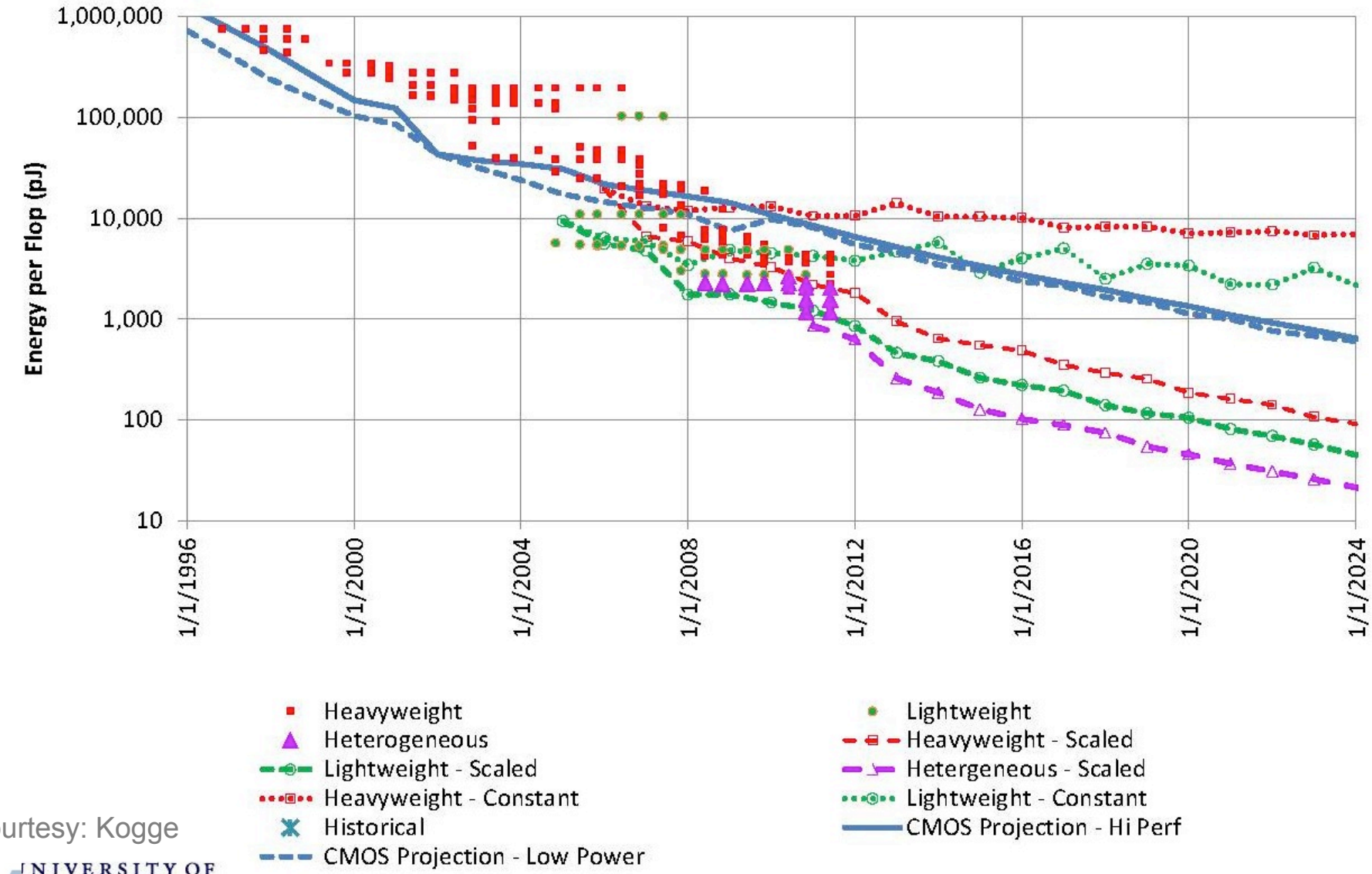
Exascale



- Goal: Reach exaflop performance in 10 years
- Why: Needed to advance knowledge in a broad range of scientific disciplines
  - Climate Science, High Energy Physics, Nuclear Physics, Fusion Energy, Nuclear Energy, Biology, Material Science and Chemistry
- Obstacles:
  - Scalability (to ~1B threads)
  - Power consumption (1/10 as compared to business as usual)
  - Resilience
  - Memory and storage bandwidth
  - Noise



Energy per Flop



Courtesy: Kogge

It's All About Communication (Kogge)

- 2015 technology:
  - 10 pJ per flop in core (10 MW per Exaflop)
  - 475 pJ per flop total (475 MW per Exaflop!)

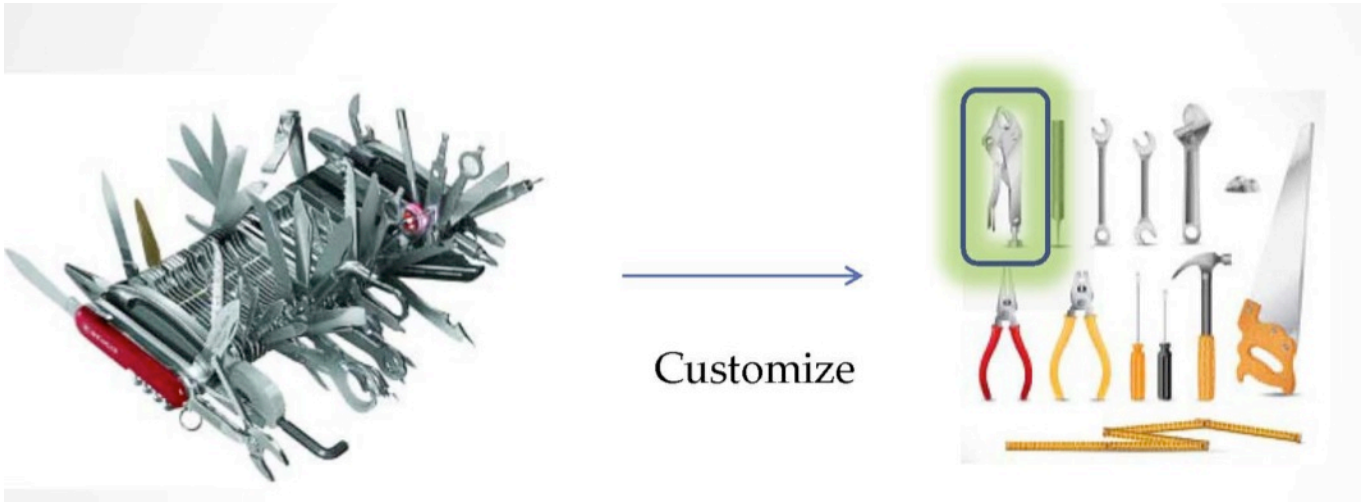
- Need to:
  - Reduce cost of communication (hw/sw)
  - Reduce communication (sw/algorithms)

Operation	Energy (pJ/bit)
Register File Access	0.16
SRAM Access	0.23
DRAM Access	1
On-chip movement	0.0187
Thru Silicon Vias (TSV)	0.011
Chip-to-Board	2
Chip-to-optical	10
Router on-chip	2

Step	Target	pJ	#Occurrances	Total pJ	% of Total
Read Alphas	Remote	13,819	4	55,276	16.5%
Read pivot row	Remote	13,819	4	55,276	16.5%
Read 1st Y[i]	Local	1,380	88	121,400	36.3%
Read Other Y[i]s	L1	39	264	10,425	3.1%
Write Y's	L1	39	352	13,900	4.2%
Flush Y's	Local	891	88	78,380	23.4%
Total				334,856	
Ave per Flop				475	

Research Directions for Exascale: Hardware

- Hardware technologies for reduced communication cost:
  - Processor, memory, interconnect
- "Black silicon": 10x10 architecture (A. Chien)
  - Use customized cores
  - Only one core used at a time



Research Directions for Exascale: Software

- OS: power management, failure management (containment, local restart)
- Runtime: user mode resource management (scheduling, memory); locality management; communication management
- Programming environment: multilevel programming; asynchronous execution; error detection (?); containment; implicit-explicit communication; refactoring
- I/O & storage: access pattern – aware storage; integration of non-volatile memory
- Algorithms: scalability; asynchrony; communication reduction

BACKUP

NOTES

- For a given EOT, MG (multigate) devices have much better gate-over-channel control than bulk/SOI devices.
- In order to compensate for the decreased gate-over-channel control for smaller gate lengths, EOT has to be reduced. Yet too small EOTs result in unacceptable **gate leakage** (tunneling) currents which also result in power loss.
- **Source-to-drain leakage current $I_{sd,leak}$** arises due to insufficient gate-over-channel control and leads to power loss and other severe adverse effects. The maximum allowed $I_{sd,leak}$ is the most fundamental technology parameter in the sense that it does not change with the gate length or other parameters and is set to 100nA/µm.
- **Supply (drain) voltage V** could also be reduced in order to decrease the power consumption and delay; however any decrease in the supply voltage would result in the corresponding decrease of the **threshold voltage V_{th}** (the gate voltage at which the device is turning on). In practice this is not permissible, since the threshold voltage is already at its limit: lowering V_{th} dramatically (exponentially) increases the source-drain leakage current.
- The main speed characteristic of high-performance logic is the **intrinsic delay τ** (the time taken by the gate to produce an output after giving the input)
- **Gate capacitance C_g** should be small in order to decrease intrinsic delay; smaller gate lengths decrease gate capacitance, yet smaller EOTs result in higher gate capacitances.
- **Drive current I_D** is thought to be the main mean to reduce intrinsic delay $\tau = CV/I$, yet its increase usually results in higher source-to-drain leakage currents.

EOT = equivalent oxide thickness
C_g = total gate capacitance
V_{th} = threshold voltage
I_D = drive current
$I_{sd,leak}$ = source-drain leakage
$\tau = CV/I$, intrinsic delay