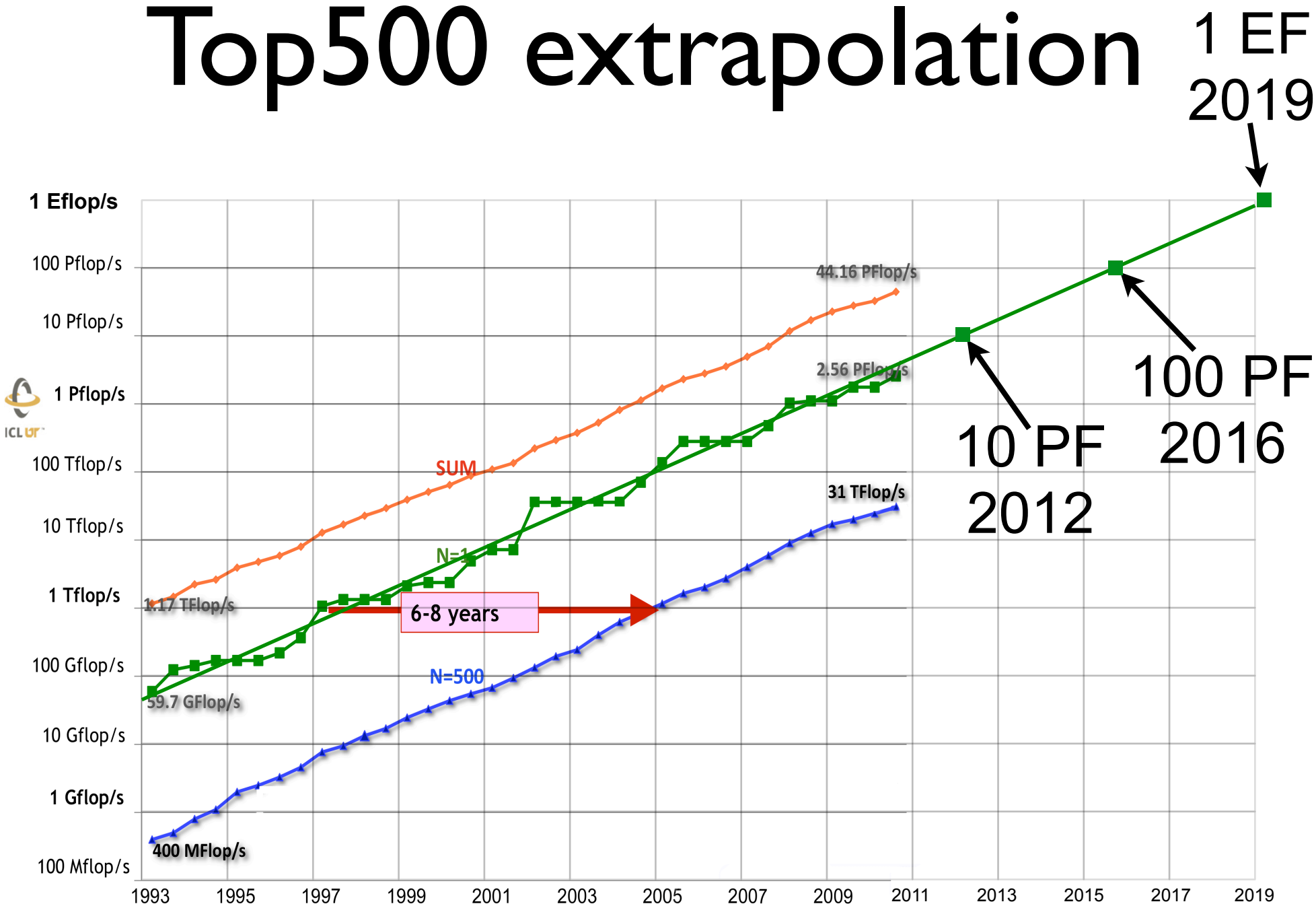
DOE ASCR 2011 Workshop on Exascale Data Management, Analysis, and Visualization

Mandate

DOE Office of Advanced Scientific Computing Research has asked the scientific data management, analysis, and visualization community to publish a joint report about recommended research as we approach scientific discovery at the exascale.

We were the *advisors* to ASCR for what research will enable this scientific discovery.

Top500 extrapolation



This workshop can have a wide impact

- Long term view: ~7 years
- Not really "SciDAC"
- DOE Office of Science base program
- Other DOE exascale computing calls
- NSF CDI, NSF TeraGrid

Day One

- Focused on information gathering and discovery. Meant to primarily spur discussions and thought.
- **Architecture talks**: What will the coming HPC architectures and infrastructure look like? What is the environment in which we will work? What constraints will we have to live under?
- **Applications talks**: What science will be explored? What challenges do we expect will be uncovered? What are the barriers to scientific understanding?
- **Technology talks**: What techniques have been successful at meeting understanding challenges to date? What technologies might be successful in breaking down the barriers? Where is further research needed? Can we identify the elements of a successful effort?

Architectures

- Andy White (LANL)
- Steve Poole (ORNL)

Sandia National Laboratories | OAK RIDGE National Laboratory | Los Alamos NATIONAL LABORATORY EST. 1943

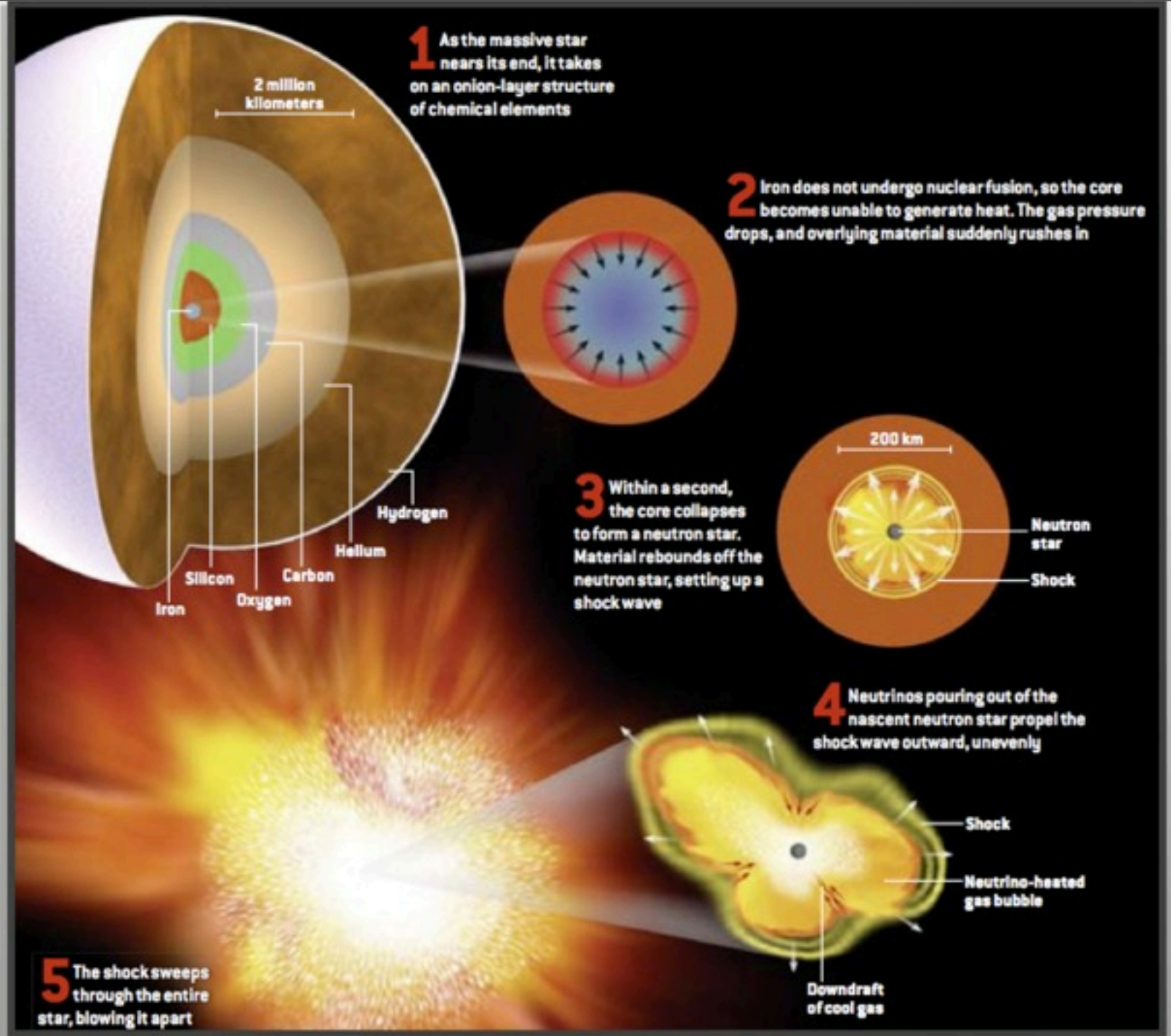
Science Partnership for Extreme-scale Computing

System architecture targets are aggressive in schedule and scope.

System attributes	2010	"2015"		"2018"	
System peak	2 PF/s	200 Petaflop/sec		**≥ 1 Exaflop/sec**	
Power	6 MW	15 MW		**≤ 20 MW**	
System memory	0.3 PB	5 PB		**64 PB**	
Node performance	125 GF/s	500 GF/s	5 TF/s	**1 TF/s**	**10 TF/s**
Node memory BW (consistent with 0.4 B/F)	25 GB/s	200 GB/s	2 TB/s	**400 GB/s**	**4 TB/s**
Node concurrency	12	100	1,000	**1,000**	**10,000**
System size (nodes)	18,700	400,000	40,000	**1,000,000**	**100,000**
Node link BW (consistent with 0.1 B/F)	1.5 GB/s	50 GB/sec	0.5 TB/sec	**100 GB/s**	**1 TB/sec**
Mean time before application failure	days	≥ 24 hours		**≥ 24 hours**	
IO	0.2 TB/s			**60 TB/s**	

Applications

- Bronson Messer (ORNL):
  Core-collapse supernovae
- Gary Strand (NCAR):
  Climate simulation
- Jackie Chen (Sandia):
  Combustion
- C-S Chang (NYU):
  Fusion



Hillebrandt & Janka 200 (Sci Am)

Data Volumes

GenASiS

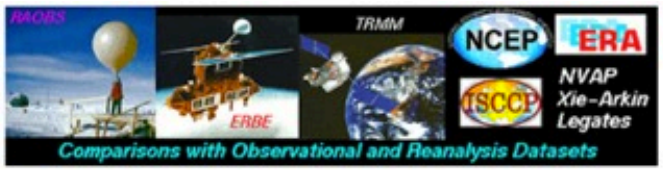
Code	# Variables	Resolution	# Dumps	Total Volume	Runtime	Machine
CHIMERA 1.0	~ 200	576X96X192	3000	~50 TB	~ 3 Months	1 PF
CHIMERA 2.0	~ 350 (expanded nuclear network α to 150-species)	576X96X192	3000	~90 TB	~ 3 Months	20 PF
GenASiS	~5000	512X512X512	3000	~30 PB	?	1 EF

- GenASiS data get large quickly as we move from moments of the distribution function to the distribution function itself
- Otherwise, data sizes increase modestly
  - The exascale machine will be a “strong scaling” platform (relative dearth of memory)

Typical viz

AMWG Diagnostics Package

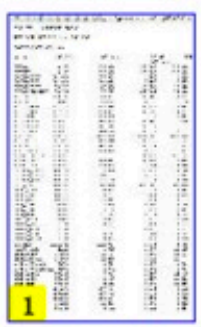
b40.1850.track1.1deg.006



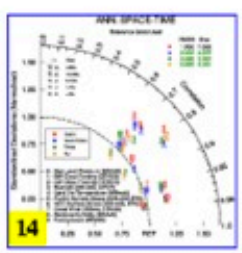
Plots Created
Wed Jan 27 19:45:17 MST 2010

Set Description

1 Tables of ANN, DJF, JJA, global and regional means and RMSE.
2 Line plots of annual implied northward transports.
3 Line plots of DJF, JJA and ANN zonal means
4 Vertical contour plots of DJF, JJA and ANN zonal means
4a Vertical (XZ) contour plots of DJF, JJA and ANN meridional means
5 Horizontal contour plots of DJF, JJA and ANN means
6 Horizontal vector plots of DJF, JJA and ANN means
7 Polar contour and vector plots of DJF, JJA and ANN means
8 Annual cycle contour plots of zonal means
9 Horizontal contour plots of DJF-JJA differences
10 Annual cycle line plots of global means
11 Pacific annual cycle, Scatter plot plots
12 Vertical profile plots from 17 selected stations
13 ISCCP cloud simulator plots
14 Taylor Diagram plots
15 Annual Cycle at Select Stations plots

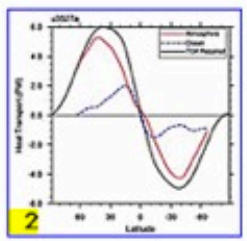


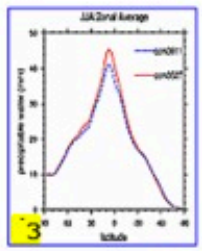
TABLES

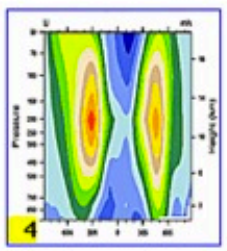


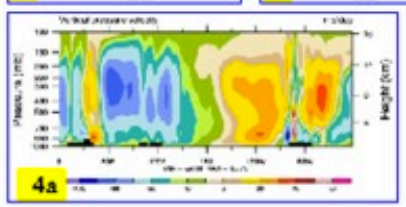
METRICS

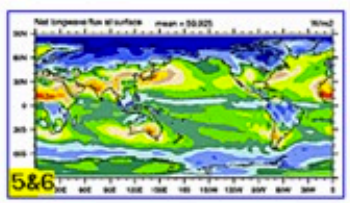
Click on Plot Type

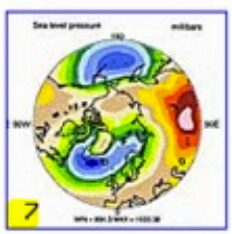


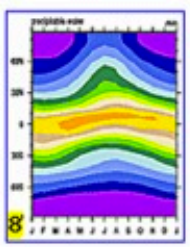


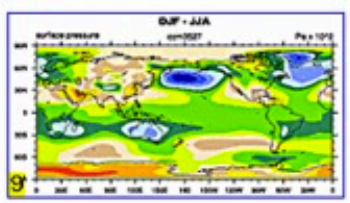


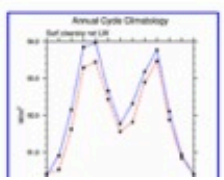


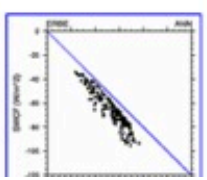


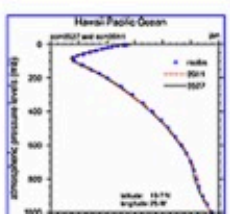


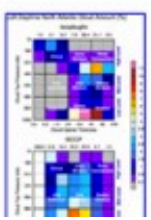




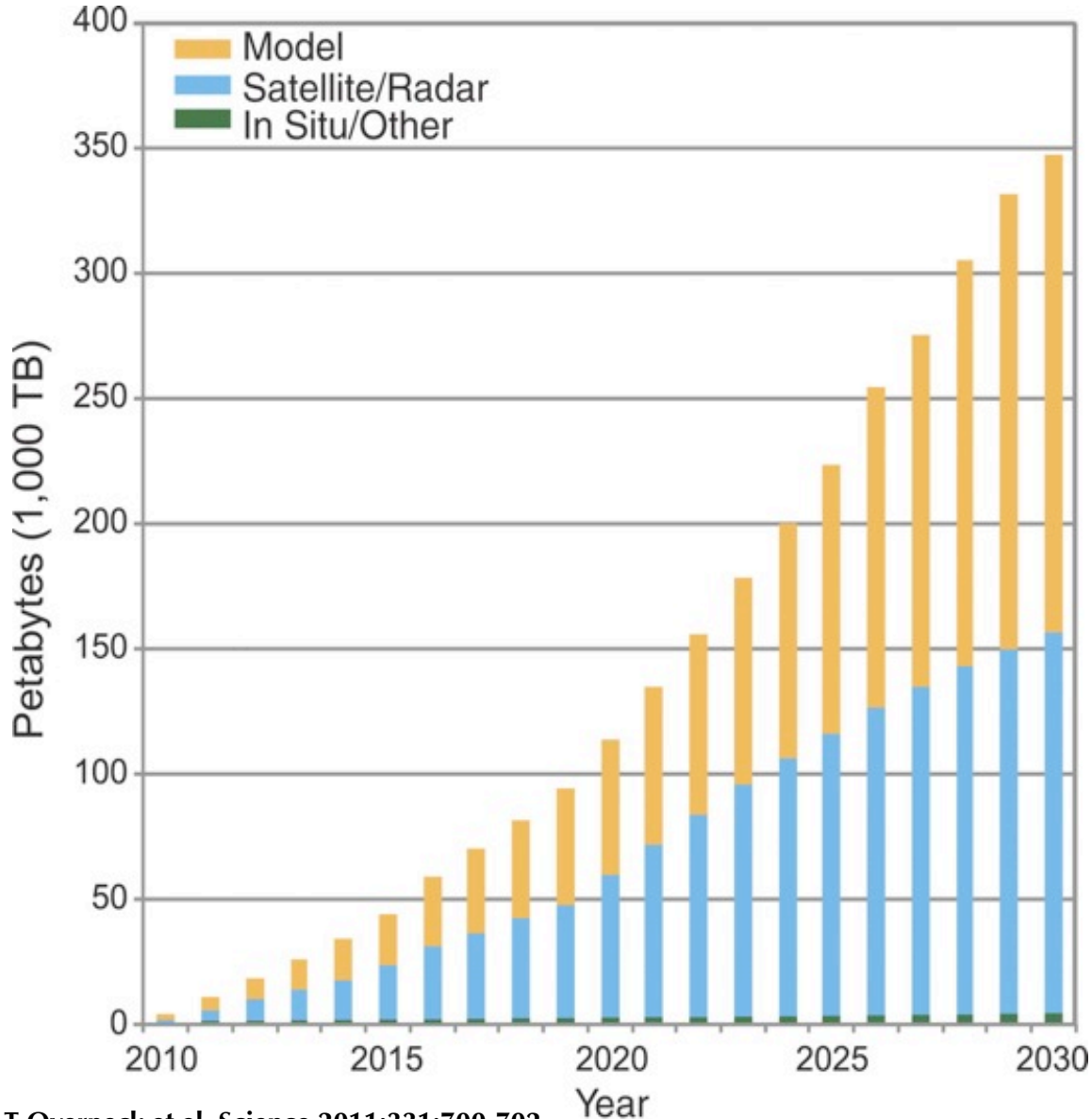








"The volume of worldwide climate data is expanding rapidly, creating challenges for both physical archiving and sharing, as well as for ease of access and finding what's needed, particularly if you are not a climate scientist."



Science
AAAS

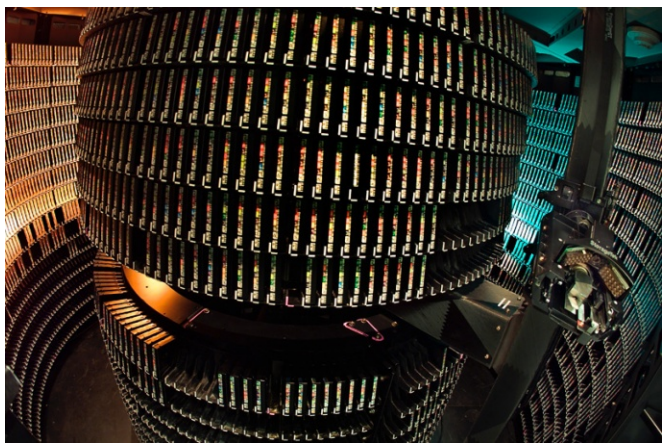
Published by AAAS

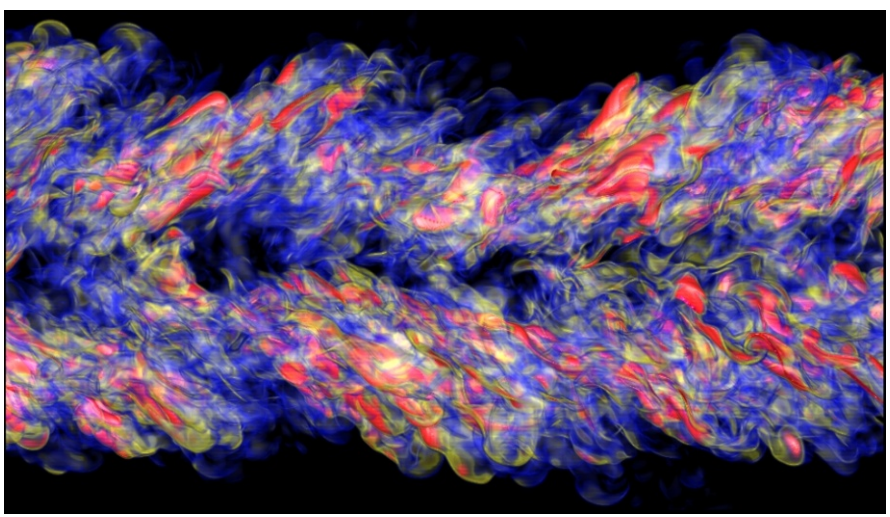
J T Overpeck et al. Science 2011;331:700-702

Challenges of Petascale DNS: Mountains of Data

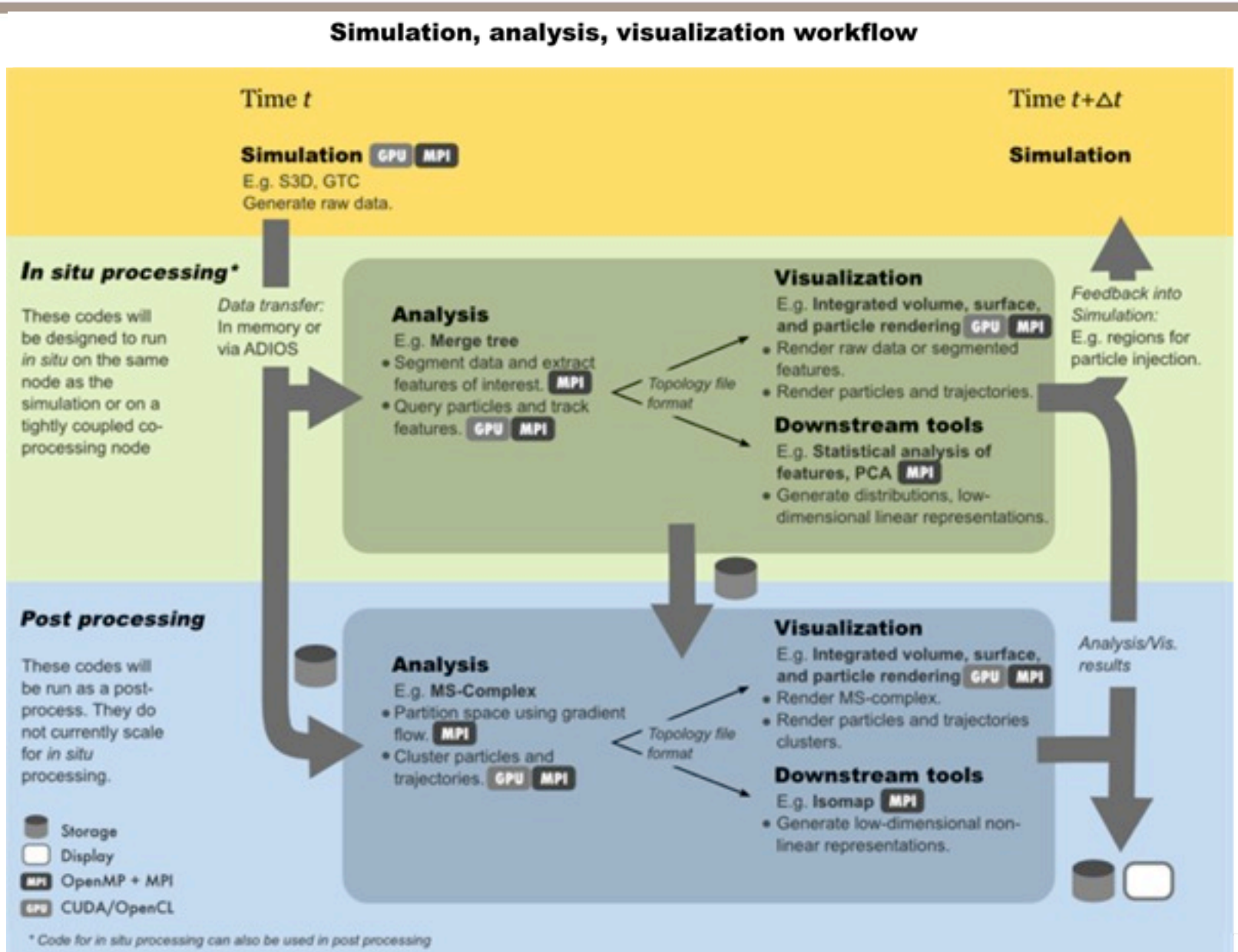
- Data size:
  - –O(3/4 PB) raw field data and 7TB of particle data on Jaguar CrayXT5, I/O 20 GB/s ADIOS
- Data complexity:
  - –Data is multi-variate (~50 species)
  - –Turbulence chaotic phenomena:
    - –Wide range of scales
    - –High intermittency, higher moments matter!
    - –Time-varying
    - – Organized coherent motions
    - –Non-locality important for spatial and temporal correlation of scalars and vectors

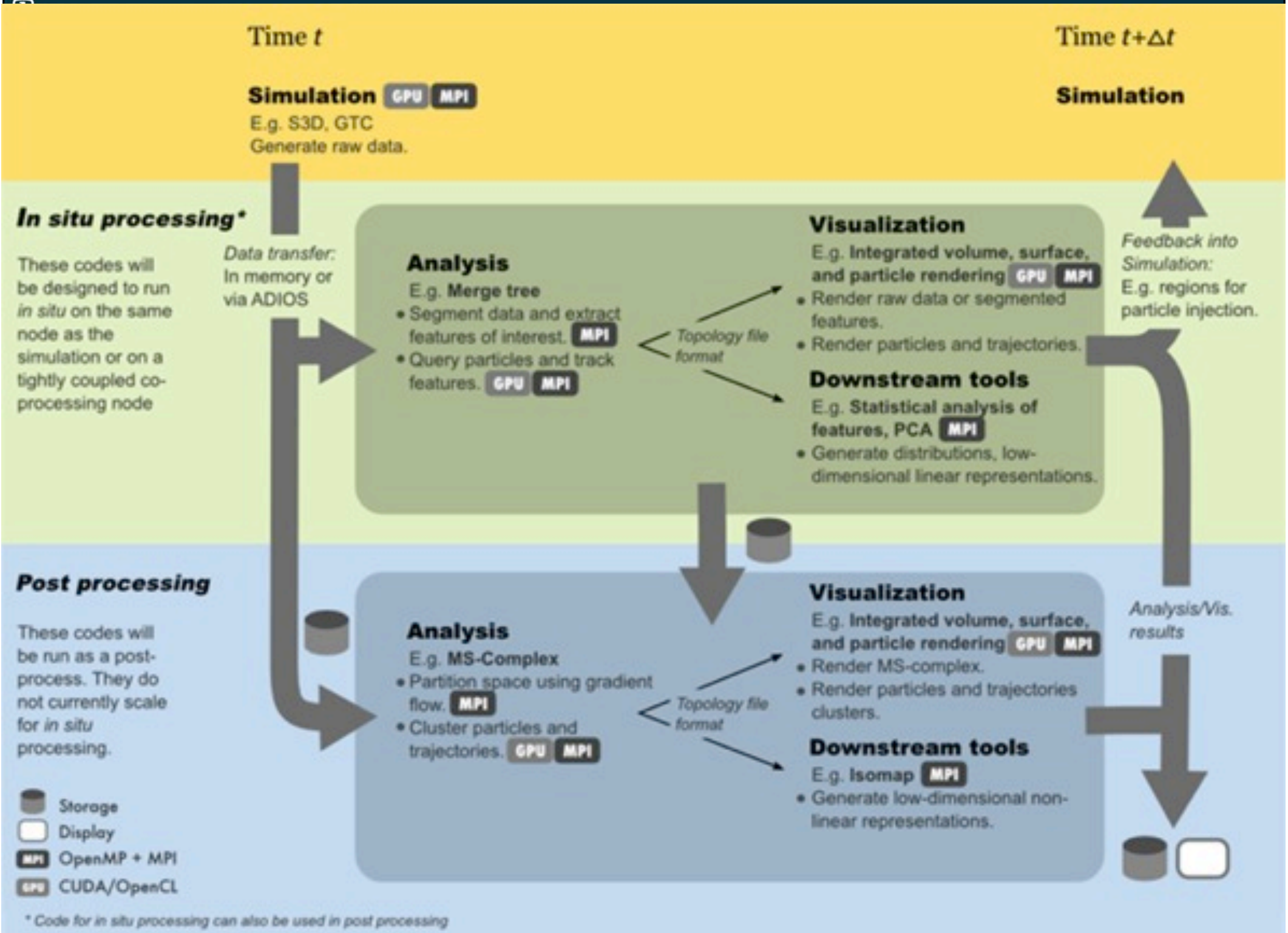
•**HPSS storage facility at NERSC**





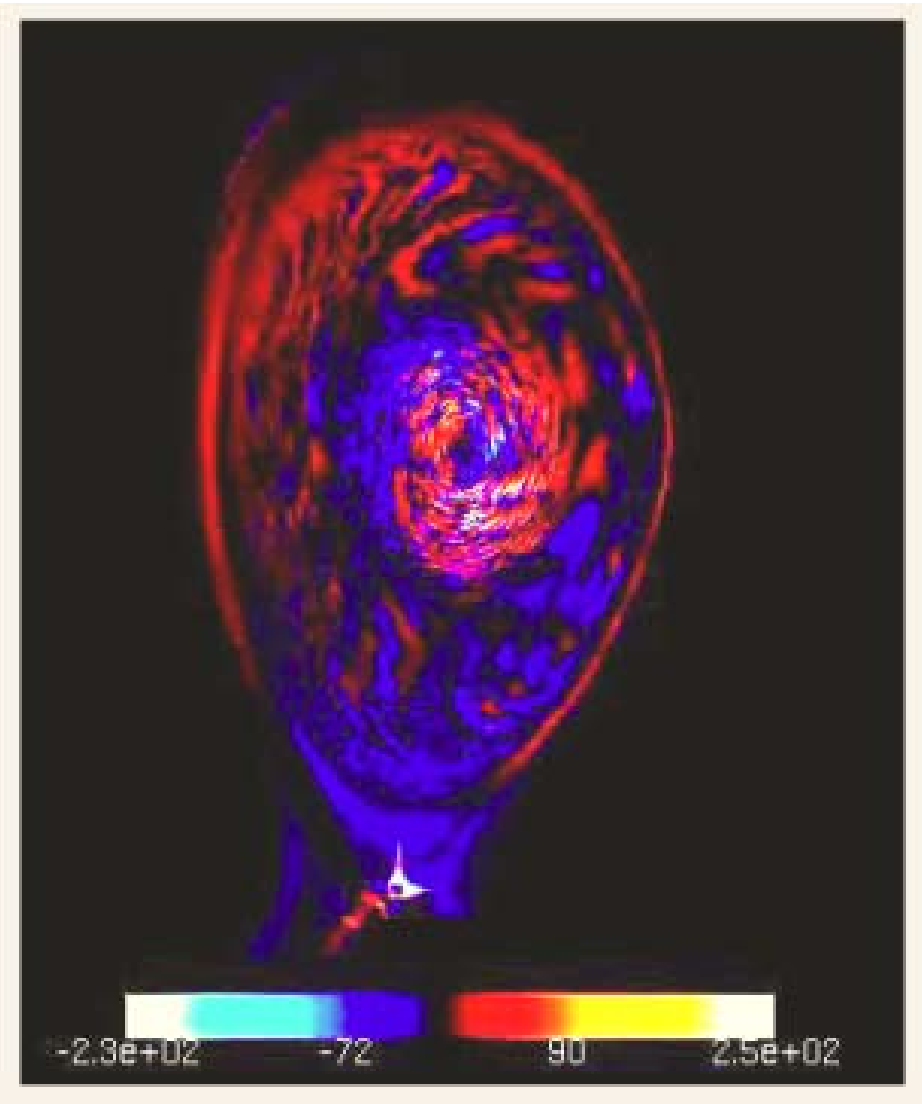
Simulation, analysis and visualization workflow : towards in situ at exascale




Time t
Time t+Δt
Simulation GPU MPI
E.g. S3D, GTC
Generate raw data.
Simulation
In situ processing*
These codes will be designed to run in situ on the same node as the simulation or on a tightly coupled co-processing node
Data transfer: In memory or via ADIOS
Analysis
E.g. Merge tree
• Segment data and extract features of interest. MPI
• Query particles and track features. GPU MPI
Topology file format
Visualization
E.g. Integrated volume, surface, and particle rendering GPU MPI
• Render raw data or segmented features.
• Render particles and trajectories.
Downstream tools
E.g. Statistical analysis of features, PCA MPI
• Generate distributions, low-dimensional linear representations.
Feedback into Simulation: E.g. regions for particle injection.
Post processing
These codes will be run as a post-process. They do not currently scale for in situ processing.
Analysis
E.g. MS-Complex
• Partition space using gradient flow. MPI
• Cluster particles and trajectories. GPU MPI
Topology file format
Visualization
E.g. Integrated volume, surface, and particle rendering GPU MPI
• Render MS-complex.
• Render particles and trajectories clusters.
Downstream tools
E.g. Isomap MPI
• Generate low-dimensional non-linear representations.
Analysis/Vis. results
Storage
Display
MPI OpenMP + MPI
GPU CUDA/OpenCL
* Code for in situ processing can also be used in post processing

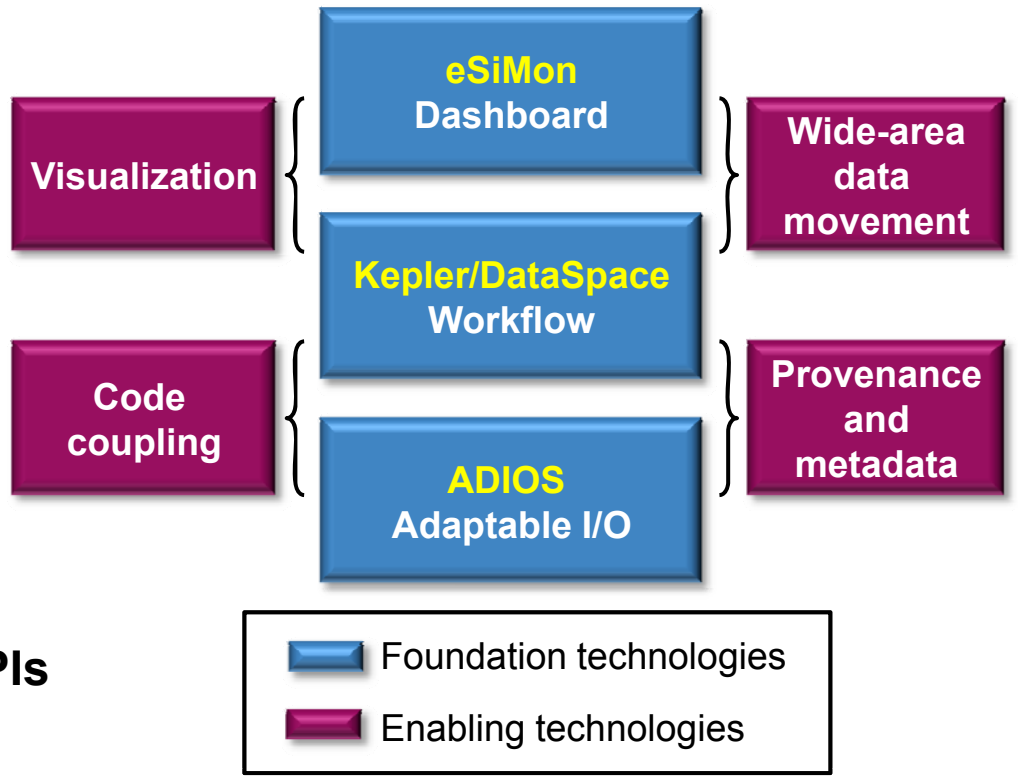
Whole-Volume, full-f ITG Simulation for DIII-D

- ITG (Ion Temperature Gradient) driven turbulence is the most robust and fundamental micro-turbulence in a tokamak plasma.
- Includes diverted edge geometry and magnetic axis
- Realistic Dirichlet BD condition Φ=0 on conducting wall.
- Heat source in the central core
- This type of simulation is possible only on extreme HPCs → needs to push the edge of future HPC
- Several new scientific discoveries have already emerged.



XGC1 deals with large scale I/O, code-coupling, analysis, and visualization

→ Our CS team responded and developed EFFIS



Need to lean only a dozen ADIO APIs

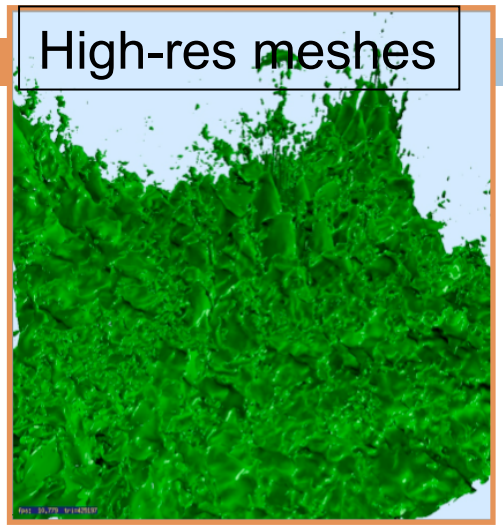
Adios_open	Adios_f_open
Adios_close	Adios_g_open
Adios_init	Adios_read_var
Adios_finalize	Adios_g_close
	Adios_f_close

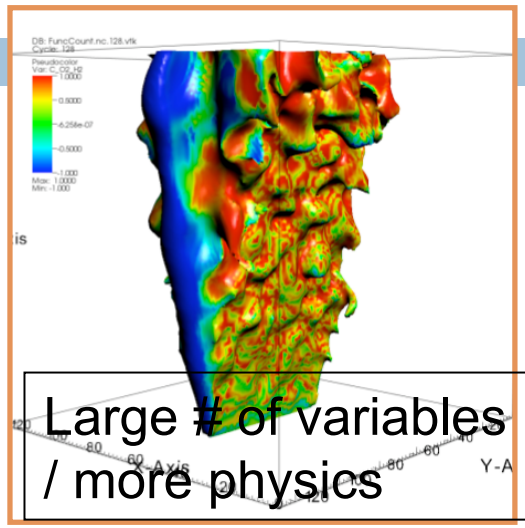
Techniques

- Hank Childs (LBNL):
  Scalable visualization
- Scott Klasky (ORNL):
  Scalable I/O & workflows
- John Wu (LBNL):
  Indexing systems
- Nagiza Samatova (NC State):
  Scalable analysis

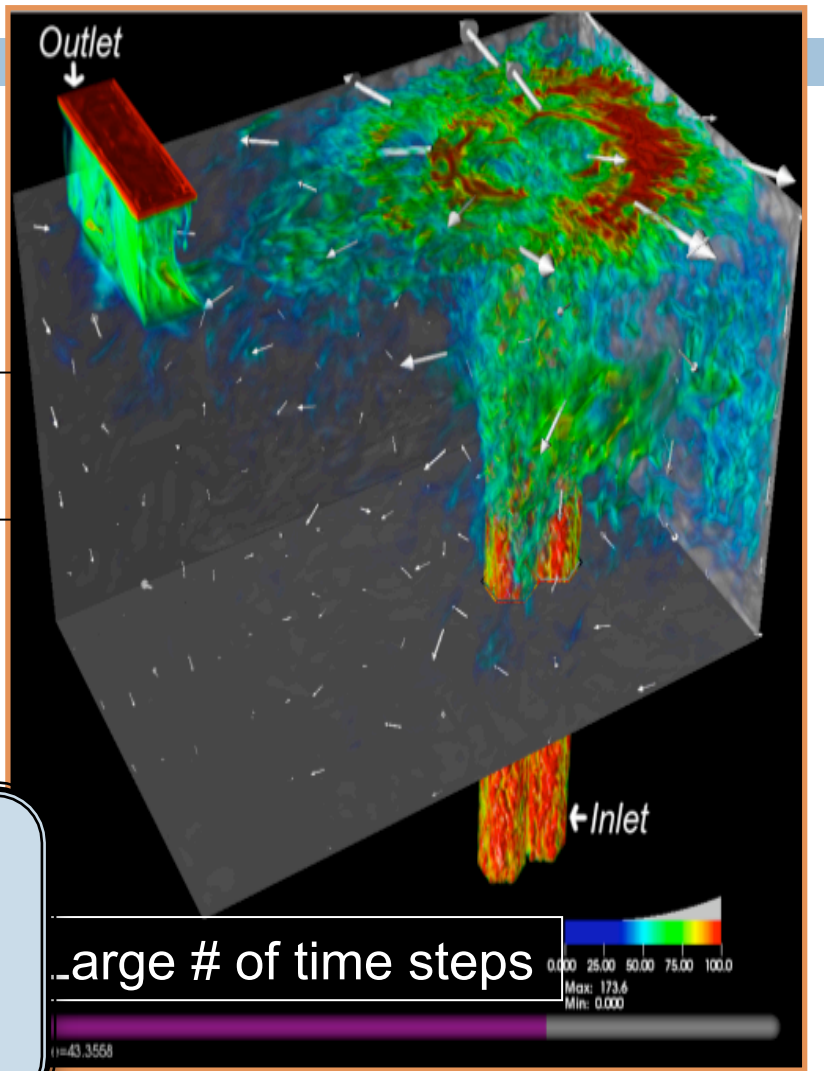
How does increased computing power affect the data to be visualized?

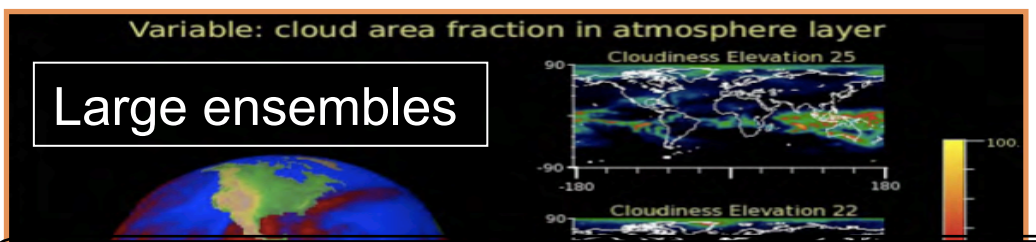
High-res meshes





Large # of variables / more physics





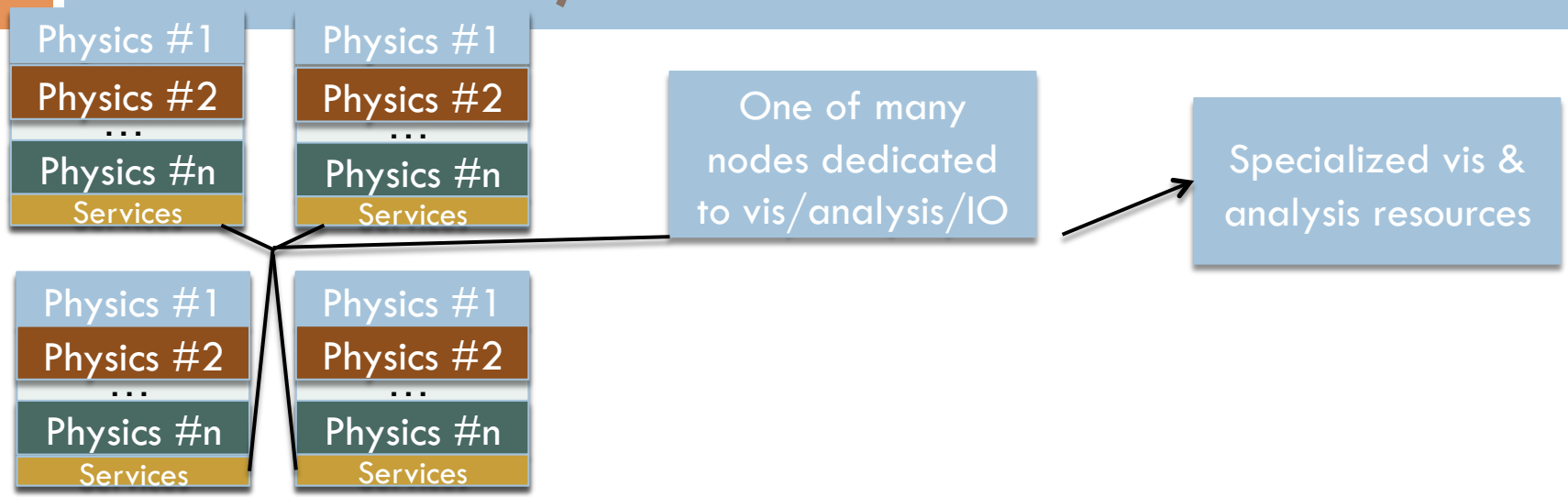
Large ensembles

Large # of time steps

Your mileage may vary; some simulations produce a lot of data and some don't.

Thanks!: Sean Ahern & Ken Joy

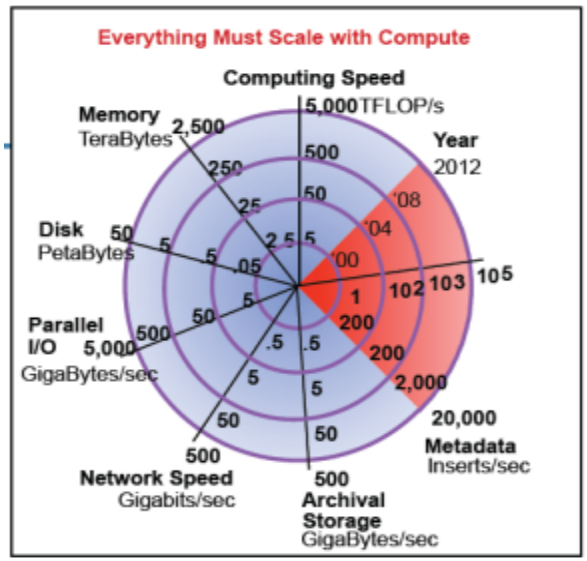
Reducing data to results (e.g. pixels or numbers) can be hard.



- Must to reduce data every step of the way.
  - Example: contour + normals + render
    - Important that you have less data in pixels than you had in cells. (*)
    - Could contouring and sending triangles be a better alternative?
  - Easier example: synthetic diagnostics

File System, Problems for the Xscale

- The I/O on a HPC system is stressed because
  - Checkpoint-restart writing
  - Analysis and visualization writing
  - Analysis and visualization reading
- Our systems are growing by 2x FLOPS/year.
- Disk Bandwidth is growing ~20%/year.
- Need the number of increase faster than the number of nodes
- As the systems grow, the MTF grows.
- As the complexity of physics increases, the analysis/viz. output grows.
- Need new and innovative approaches in the field to cope with this problem.
- The biggest problem is the $$$ of I/O, since it's not FLOPS



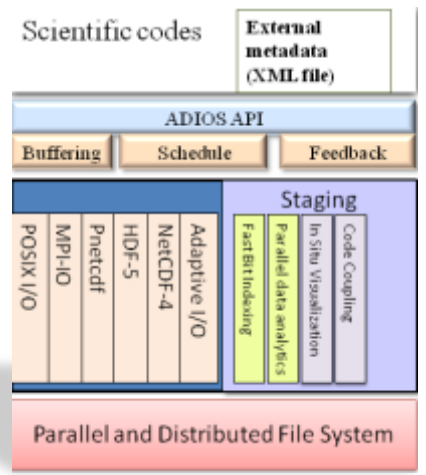
Garth Gibson 2010

Research to look at for the exascale

- I/O pipelines need be constructed to include
  - Indexing.
  - Analytics.
  - Multi resolution analysis.
  - Writing to the file system.
  - Compression.
- Different I/O methods for different I/O patterns for both reading and writing. (Restarts, Analysis, Visualization, …).
  - I/O for Analysis and visualization need to be re-examined for multi-resolution.
- Checkpoint in memory first, and then stage to disk, if necessary.
- Rethink file formats for resiliency.
- Must think carefully about QoS-like techniques.
- Helpful to include I/O componentization.
- Rethink for energy cost/savings.

Our solution:
ADIOS: Adaptable I/O System

- Provides portable, fast, scalable, easy-to-use, metadata rich output
- Simple API
- Change I/O method by changing XML file only
- Layered software architecture:
  - Allows plug-ins for different I/O implementations
  - Abstracts the API from the method used for I/O
  - New file format (ADIOS-BP), for petascale- exascale.
- Open source:
  - http://www.nccs.gov/user-support/center-projects/adios/
- Research methods from many groups:
  - Rutgers: DataSpaces/DART
  - Georgia Tech: DataTap
  - Sandia: NSSI, Netcdf-4
  - ORNL: MPI_AMR

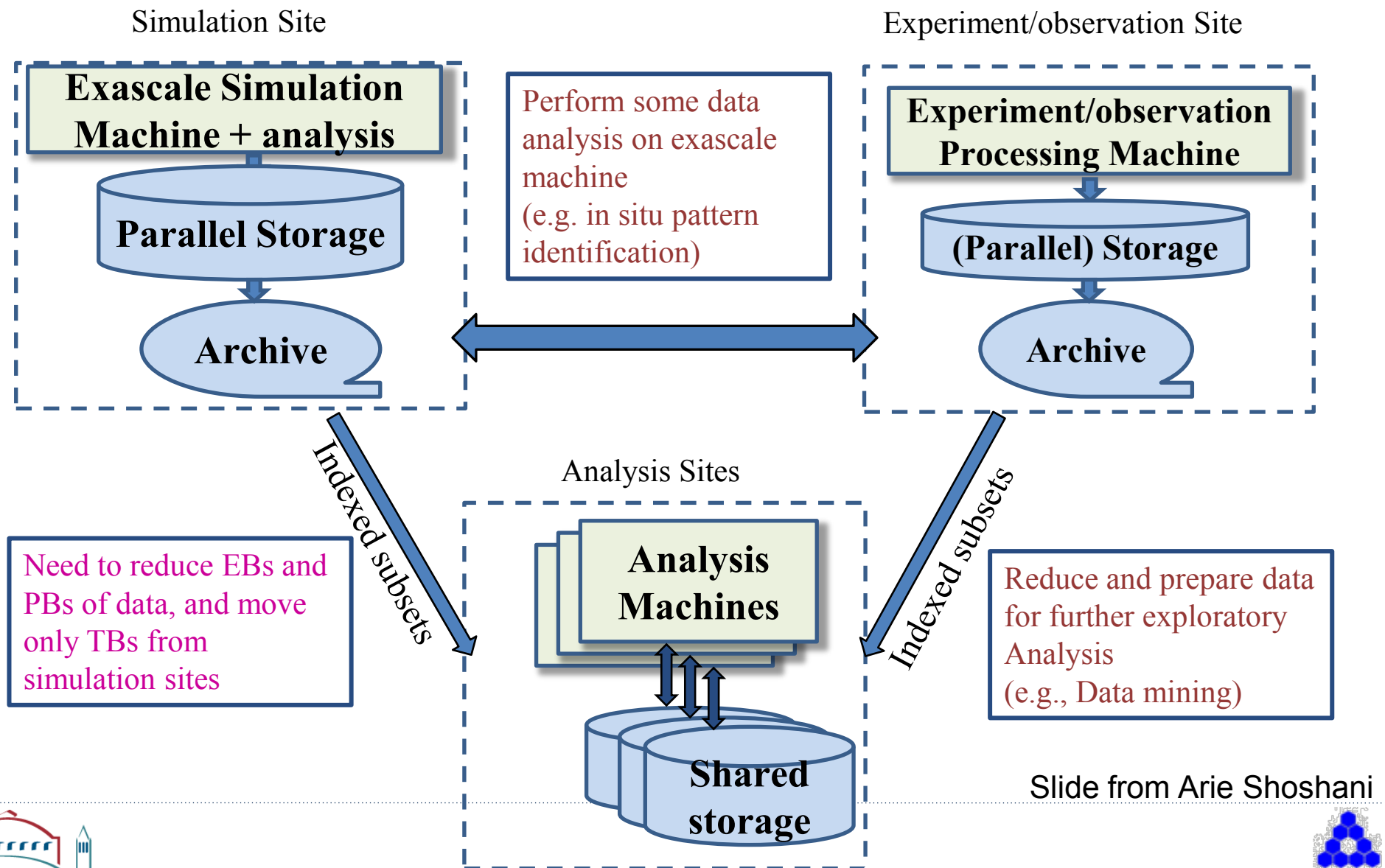


Indexing Can Help: Example with Particle Tracking

- Collaboration between SDM and VACET centers
  - Query driven visualization
- Use FastBit indexes to select and track the most interesting particles
  - 100 – 1000 X faster than brute-force approaches

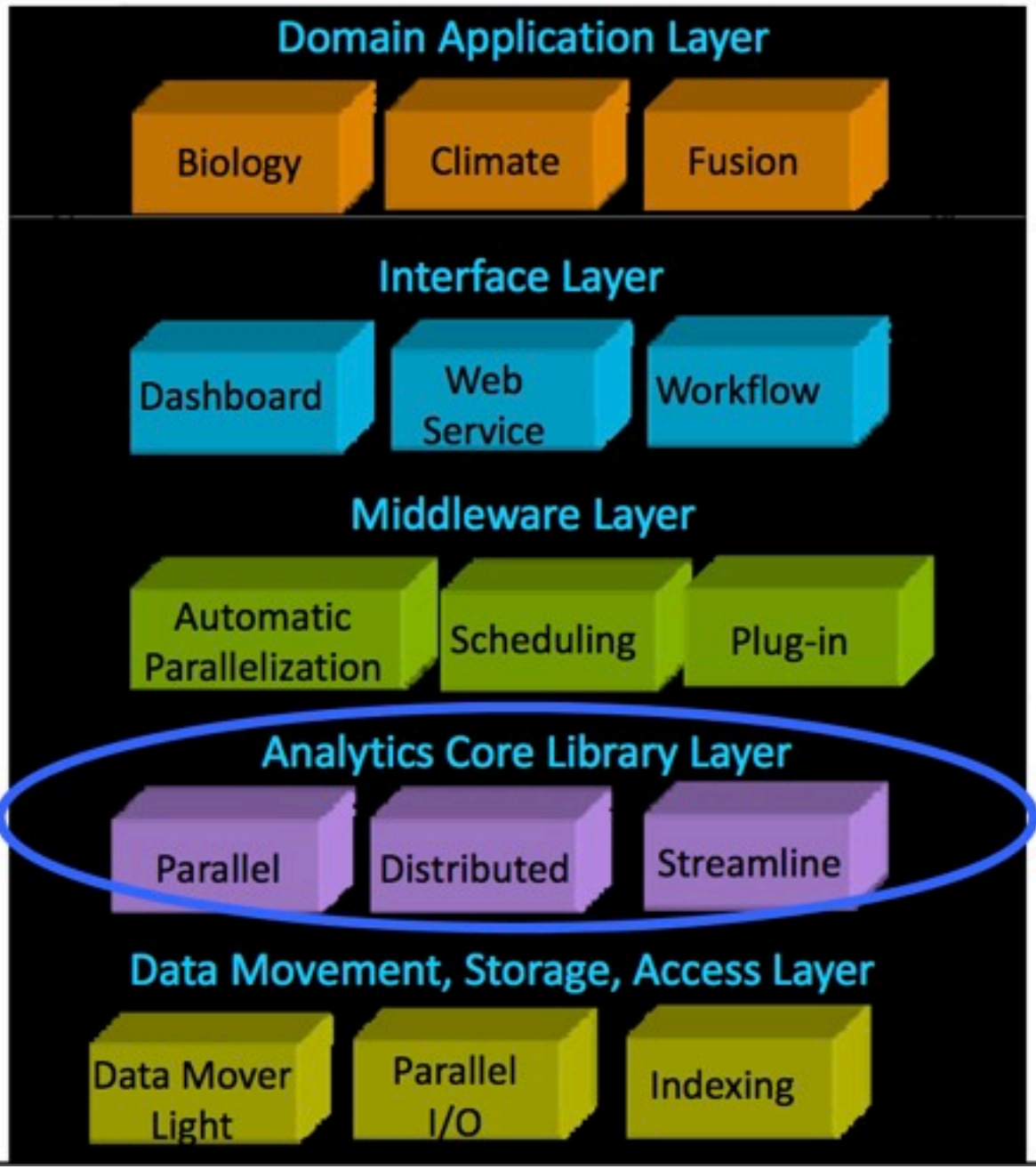
Ruebel et al SC08

Data Access Challenges: Indexing is Only a Small Part of the Story



Slide from Arie Shoshani

End-to-End Data Analytics Software Stack is Complex: Generic (All Applications) Perspective



Exemplars

CDAT

Kepler, VisIt, SciRun, FIESTA

pR, VisIt

RScaLAPACK, NW-Minebench

FastBit, ADIOS pnetCDF, HDF5

Focus of my talk

OAK RIDGE National Laboratory

UNIVERSITY Computer Science

In Situ Preconditioners for Scientific Data Compression

- **Myth: "Scientific data is almost uncompressible."**

GTS Fusion Simulation Data (Stephane, PPPL)

C&R Data

- ~2TB per C&R
- Every 1 hour
- Two copies
- Keep the last copy

Analysis Data

- ~2TB per run (now)
- Every 10th time step
- Cannot afford storing all b/s of
- Analysis routines and I/O reads
- Matlab analysis routines

V&V Data

- Small
- Every 2nd time step

Expected: 10-fold increase by 2012-2014

Reports to read

- 2007 Salt Lake City report
- Ten application exascale reports
  - ASCAC summary report
- Two reports in email earlier today
  - Cross-cutting workshop report (CS and Math issues)
  - Dec `09 architectures workshop report
- Nagiza's summary slides – Sent out in email this evening

Day Two

- The real work begins.
- With architectures in mind, scientific applications generating data, and technologies that can be brought to bear…
  - What research topics are likely to be fruitful in meeting the challenges?
  - How will we measure success?
  - What will be the expected fruits?

Questions to address

- What are the research challenges to enable science at the exascale (vis, sdm, analysis, realtime/in situ, post processing)?
  - What are the programming and data models for exploiting exascale architectures?
  - How do we consider power efficiency in our analysis algorithms?
- Document how hardware changes will affect vis/analysis/IO
- Document impacts of research, document impacts on application science
- Metrics for success?
- What is the roadmap for research?
- Data fusion?

First breakout group session

- Concurrent processing / in situ
- I/O and storage
- Impact of exascale architectures on data post-processing

First breakout group session

- A: Concurrent processing / in situ: Donatello Room (Kwan-Liu Ma)
  - Reduction and analysis running concurrently with an application
  - Could share a node, could be co-resident on an HPC system
  - Memory hierarchies
  - Annotation of data in flight
- B: I/O and storage: Salon 11 (Scott Klasky)
  - Scientific data formats, database technologies, hierarchical storage
  - Common data models
  - Indexing, reordering, acceleration mechanisms, compression
- C: Impact of exascale architectures on data post-processing: Salon 12 (Alok Choudary)
  - Constrained memory footprint
  - NVRAM
  - I/O wall, limited I/O bandwidth
  - Power considerations of communication, reduced communication algorithms
  - GPU, accelerator technologies

Second breakout group session

- Visualization and Analysis
- Scientific Data Management

Second breakout group session

- D: Visualization and Analysis (Valerio Pascucci)
  - Multiresolution, sampling methods
  - Streaming / out-of-core algorithms
  - Correlations, integration with observational data
  - Feature tracking
- E: Scientific Data Management (Terence Critchlow)
  - Workflow systems, provenance
  - Management of ensembles, parameter studies
  - Data fusion
  - Machine learning
  - Informatics, Information visualization

Day Two: Writing

- The primary output of this workshop will be a report. We will generate the skeleton of the report here and flesh out that skeleton over the weeks following the workshop.
- Each breakout group is responsible for generating a framework describing its focus area, including major challenges, recommendations for research, and expected benefits.
- We will reconvene after writing, allowing each breakout group to present its findings to all attendees.
- We will conclude with a fusion of all three breakout group reports into a larger (rough) framework.

Current work

- The SciDAC-3 call came out while we were in the last day
- Current writing efforts
  - We produced a very rough outline at the workshop
  - We produced a slightly more detailed outline through teleconferences
  - Writing will commence in earnest in a few weeks once many deadlines are over
- Expected deadline for full report is in the late summer to early fall