Data Intensive Computing: Interacting with High-End Computing Storage Systems

David R. White
Data Analysis and Visualization
Sandia National Laboratories

Maui
March 7, 2006

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

Models of Interaction with HEC storage

- First, what is HEC? (Google: HEC)
  - Harken Energy Corp.
  - Hydrologic Engineering Center
  - Higher Education Commission of Islamabad
  - Higher Education Center for Alcohol and Other Drug Prevention
  - ….

Models of Interaction

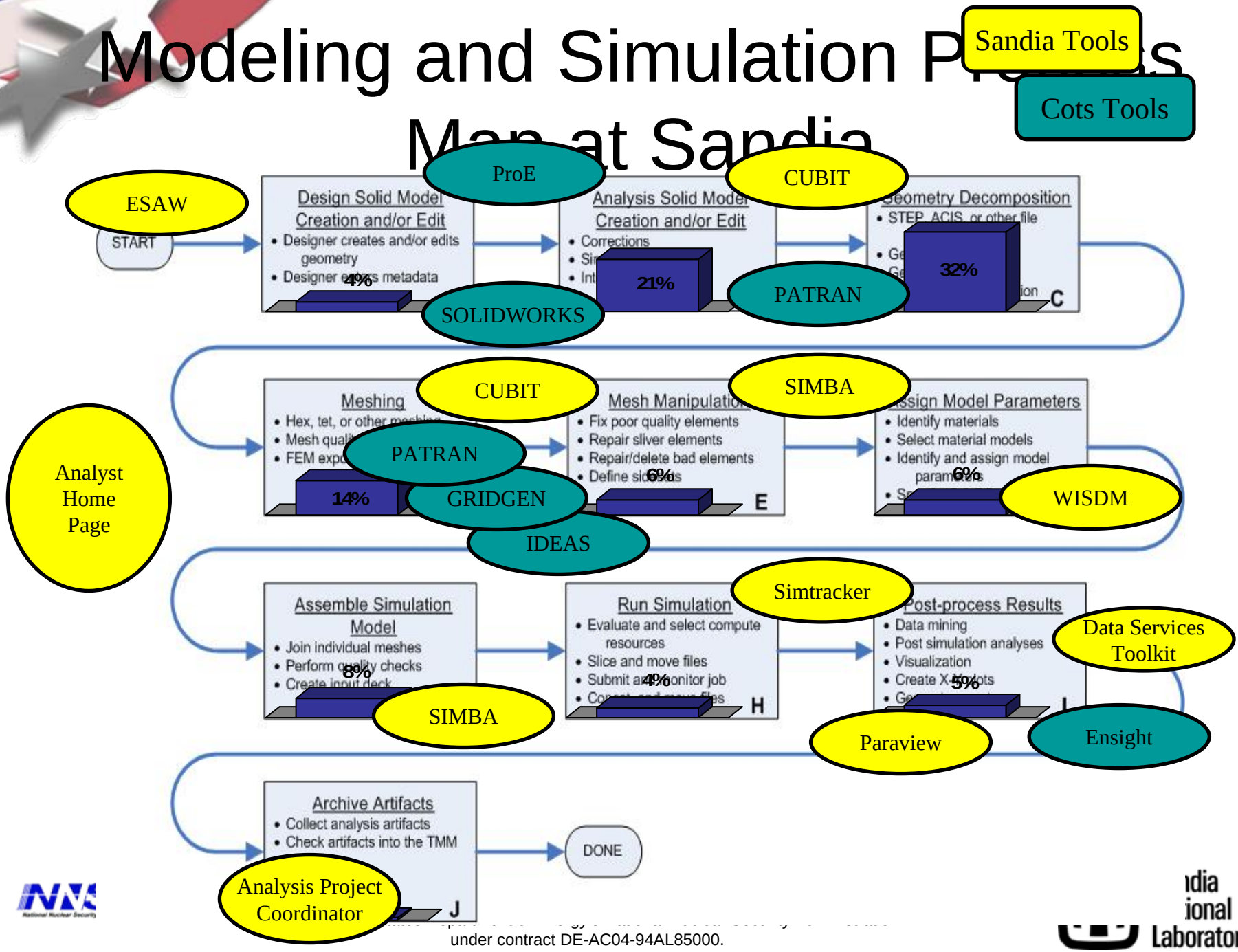
- Storage System per HEC system
  - Files and results on various platforms scattered across the complex.
    - ParaView (Parallel Vis Tool) is installed on 13 different Linux clusters around Sandia to support data visualization of computations
    - File movement means death to work flow
    - After people do ‘work’, they move it to central location (tape!)
  - Currently users must track results (files) manually
    - Results in users choosing one/two platforms to do their work meaning poor load balancing and restrictions on job sizes due to choice of system.
    - This is also do to system specific “hoops”.

HEC

- Abstracting such issues from users is critical to making HEC pervasive in engineering (or science in general)
  - If we don't, we may be working for Harken Energy Corp or Higher Education Center in Islamabad…
- But this isn't the only issue, it requires a total system solution…
  - HEC must remember this perspective…

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

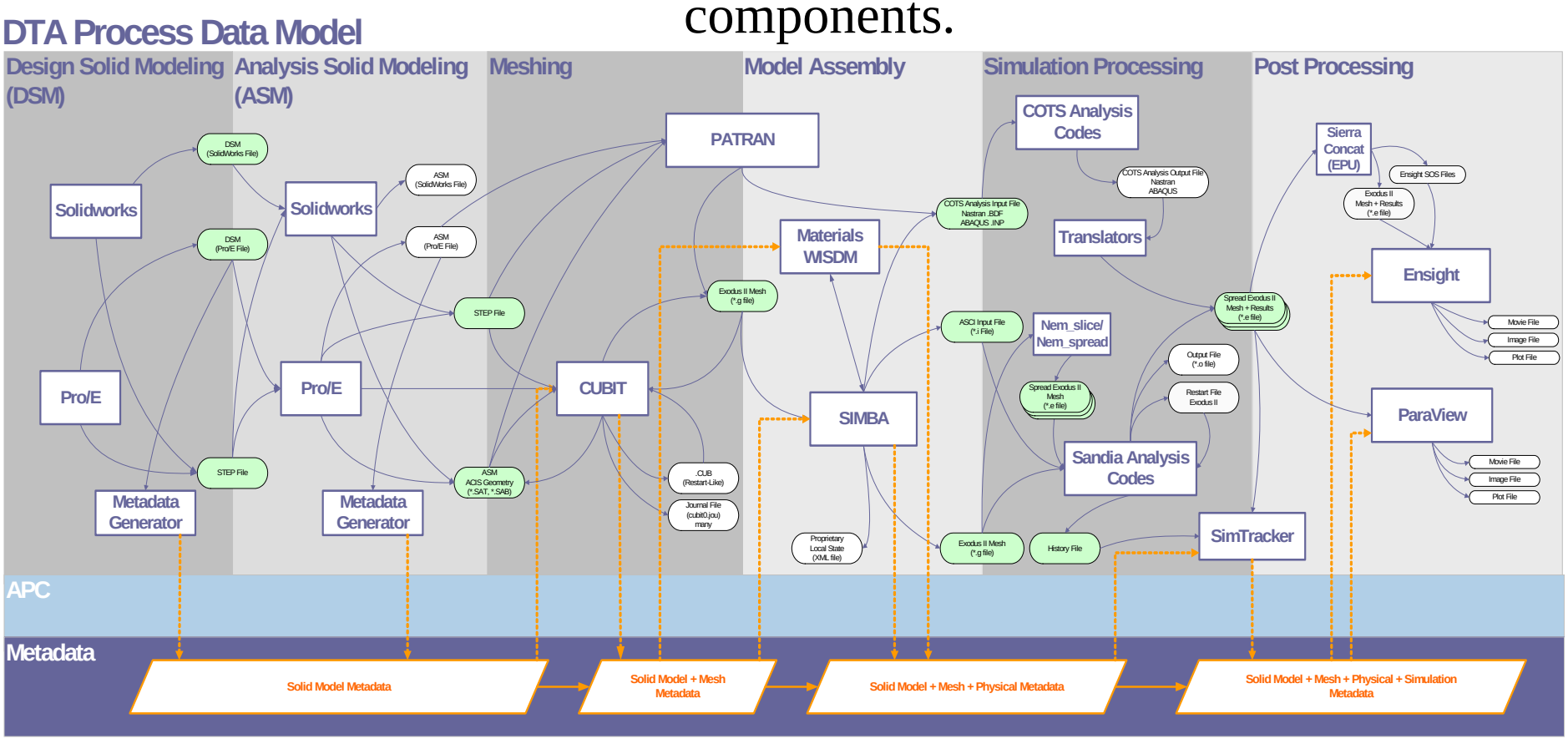
Modeling and Simulation Process Map at Sandia



under contract DE-AC04-94AL85000.

Strategy 2: Federated Integration Via Metadata

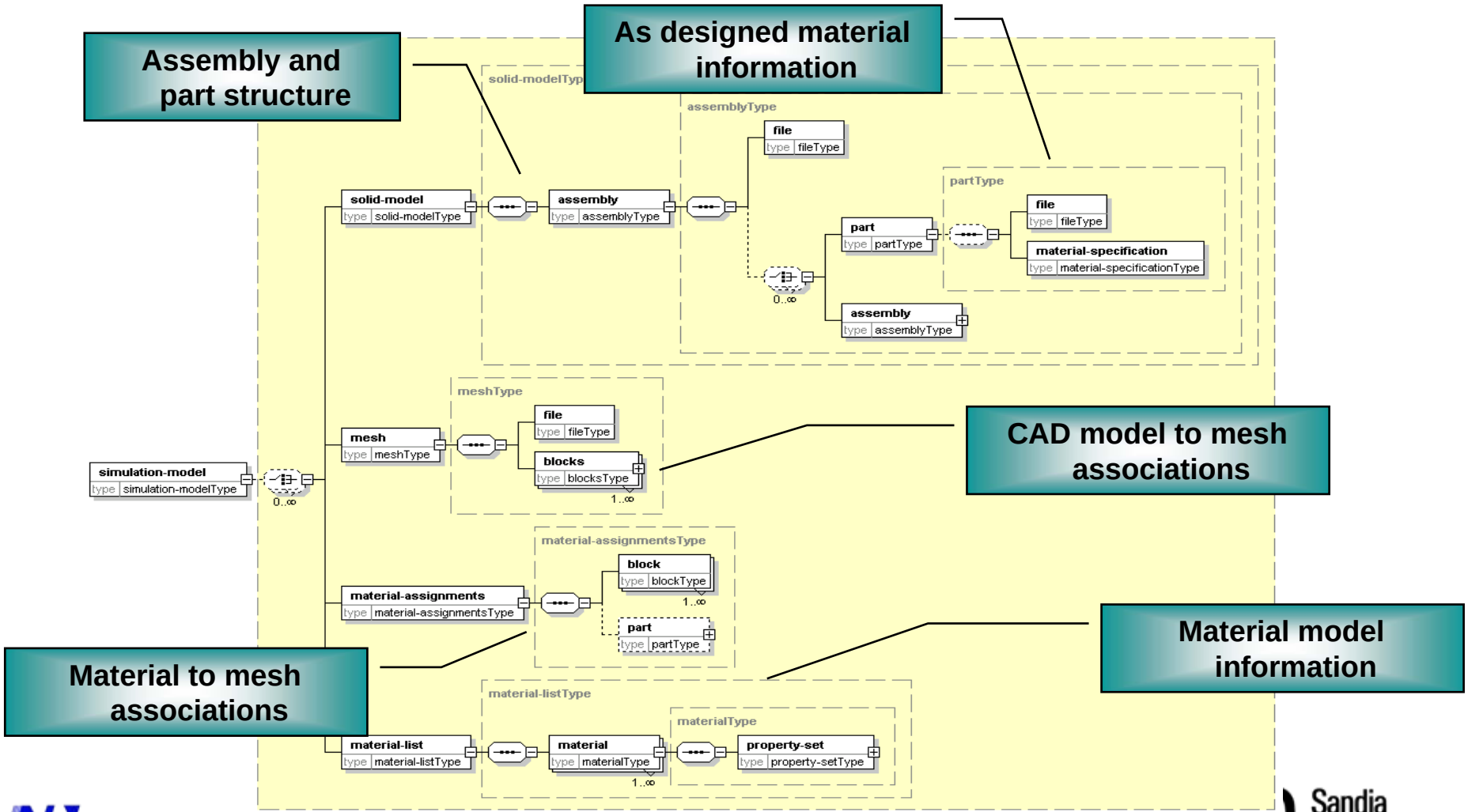
Open Architecture: A software architecture based on well defined interfaces between the core (generalized) software components.



Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

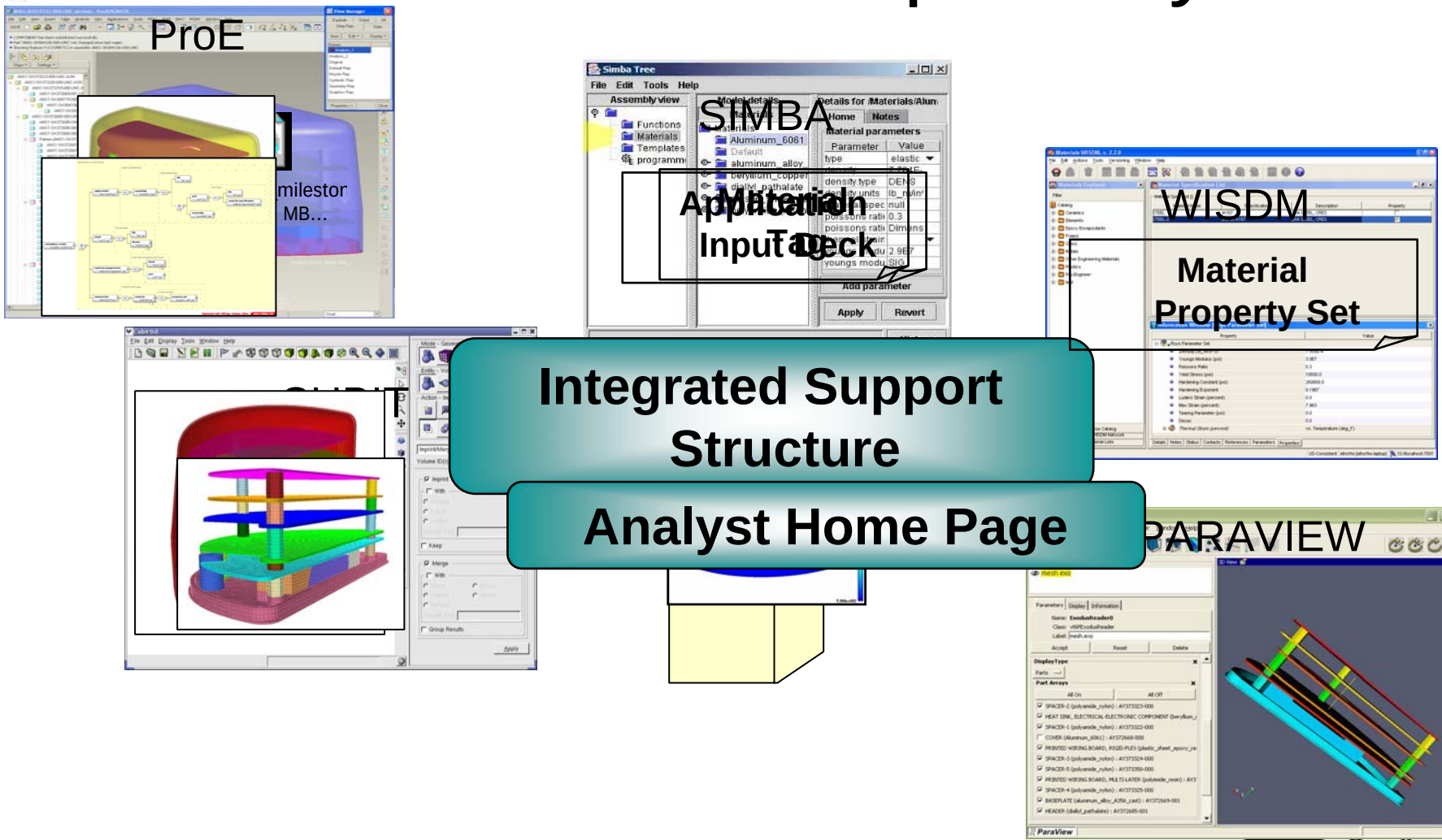
Simplified XML Metadata Schema



for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

Sandia National Laboratories

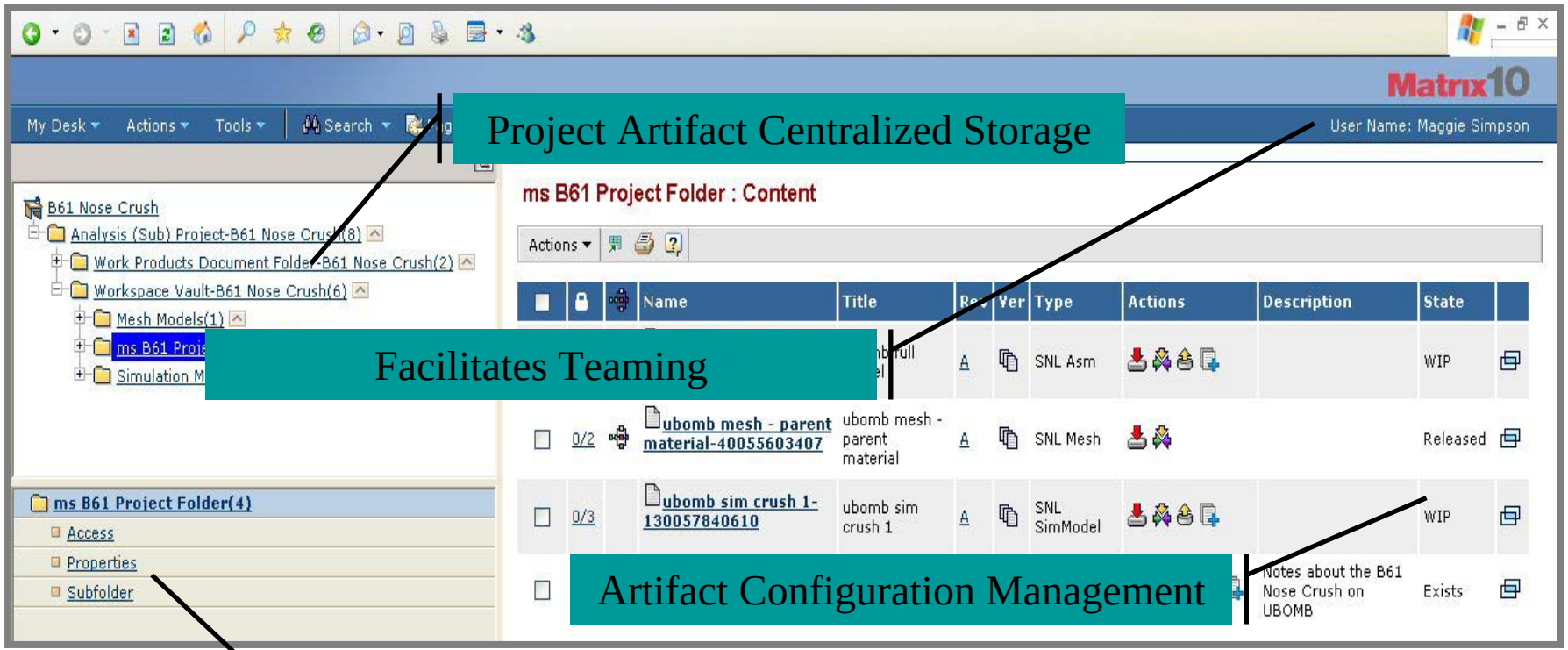
Metadata Application: Material Transparency



Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Strategy 3: Common Operating Environment

APC (Analysis Process Coordinator)



Project Artifact Centralized Storage

Facilitates Teaming

Artifact Configuration Management

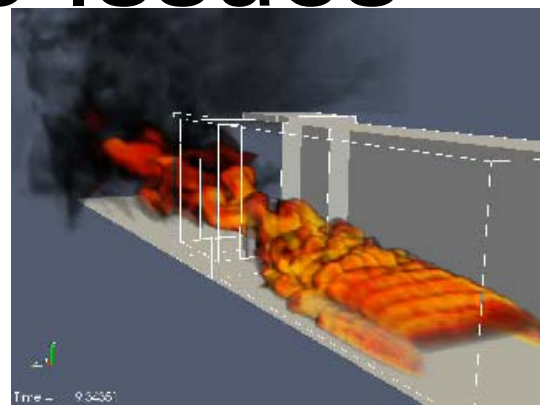
Artifact Relationship Maintenance

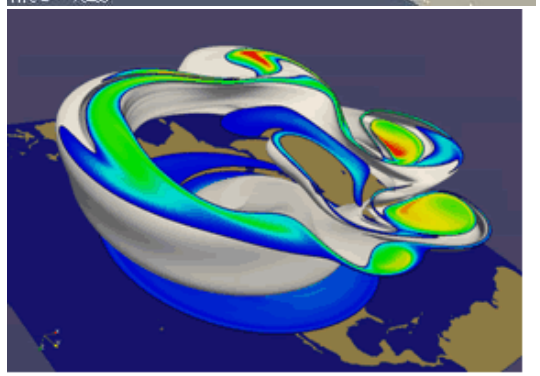
Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Simulation Coordination

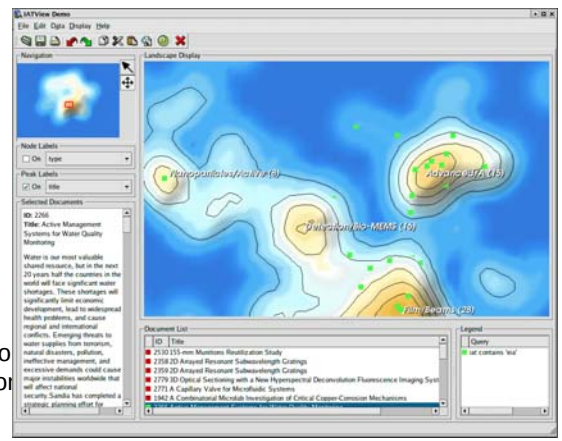
- Provide end-users abstraction so that they will run on the best machine for their particular job with-out regard to file tracking and platform specific syntax knowledge
- Ability to develop database of simulations with qualifications on detail
  - Really begins the day of data explosion where results are searchable (perhaps not exact sim results but metadata of sim) and makes HEC useful in engineering design and discovery.

Other HEC I/O Storage Issues





- "Run-time" visualization needs to be supported
  - Application steering is a misnomer for engineering labs with initial value problems
  - However, stop the calculation before it wastes 3 days or even 8 hours is very valuable.
  - This requires reading the restart files on the file system that they were written to for the computation (file transfer kills efficiency in this and may take as long as the job!)
    - So systems must have 'vis' or interactive nodes, or other systems must have access to the file system (you choose, doesn't really matter to end user)
- File systems must be designed for LOTS of small files
  - In past year, visualization switched from reading really big files to reading exact output from analysis software.
    - files created small, remain small…
- As end-to-end systems are created: (i.e. database of engineering problems) it must be storable and searchable.
  - Trend analysis, abstract inspection of results (look at all 3 million engineering results, etc…)



Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Co
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

Beyond Modeling and Simulation

- Think in terms of Petabytes/hour or N-bytes / time, coming from sensors, computing, etc… (Intelligence Community)
- Understanding IS the perspective (remember this is why we do computing…)
- Natural Language Queries/Interfaces, Interaction, Abstract views, are the future.
- I/O and storage have to enable these
  - Think 9/11 report and issues with disparate unconnected databases (politics MAY be the minor issue!!)