National Energy Research Scientific Computing Center (NERSC)

Science Driven Analytics

Wes Bethel, Cristina Siegerist, Cecilia Aragon
Visualization Group, LBNL
02 May 2006
Science Driven Analytics

- Simulations and experiments are generating data faster than it can be analyzed and understood.
- Science bottleneck: information analysis and understanding.
What is Analytics?

• **Science of reasoning.**
  – Insight and understanding from large, complex, disparate, conflicting data.

• **Visual Analytics**
  – Science of reasoning facilitated by visual interfaces.

• **Why at NERSC?**
  – Data, data and more data.
Distributed Analytics Workflow serves an entire community.

- Images collected from NEAT (Near-Earth Asteroid Tracking) telescopes.
- Images sent from telescope to network via custom wireless network.
- Images sent to NERSC for analysis on PDSF. Digital processing (registration, differencing) to locate potential targets.
- Potential Type 1a supernovae targets identified and broadcast to observation community (24-hour turnaround).
Iterative, Query-Driven Analytics Example

• Combustion research.
  – Compare simulation with experiment.
  – Need analysis in regions defined by data- and topologically defined features.
  – 10s of TB of simulation data, but only small portions interesting for any given analysis problem.
  – Several NERSC projects: Bell (LBNL), Chen (SNL-CA).
  – Need for remote analytics capability.

(Data courtesy M. Day, J. Grcar, and J. Bell, LBNL)
Analytics Challenges

• Cellular simulation: integrate and analyze data from multiple sources: proteins, multimolecular assemblies, metabolic pathways, and scale codes to complexity of living organisms.

• Fusion: analysis and comparison of multiple-code simulations and experiments leads to tokomak plasmas with improved energy confinement, and that are predictable and repeatable.

• Astrophyics: CMB/Planck – satellite mission to collect data; size estimated at 100s TB per year; needs to be stored and analyzed; workflow serves a community of over 80 researchers.
What is Analytics, Really?

• Intersection of:
  – Visualization, analysis, scientific data management, human-computer interfaces, cognitive science, statistical analysis, reasoning, ...

• No such thing as Microsoft Analytics v1.0.

• Solutions are domain-specific combinations of above technologies.
Why Analytics, Really?

- All sciences need to find, access, and store and understand information.
- In some sciences the data management (and analysis) challenge already exceeds the compute-power challenge in its needed resources.
- The ability to tame a tidal wave of information will distinguish the most successful scientific, commercial, and national security endeavors.
- It is the limiting or the enabling factor for a wide range of sciences.
Why Analytics at NERSC?

• Clear scientific need (next slide).
• Small change in program focus likely to have profound positive impact on science.
• High likelihood of success:
  – World-class support for computational science projects.
  – Excellent existing infrastructure and program.
  – Building upon well-established visualization program at NERSC.
  – Clear, focused analytics strategy.
• Users highly value institutional visualization support.
• Establish a coherent program that focuses on remote visualization.
• Establish mechanisms whereby generally applicable visualization technology is developed and deployed in a centralized fashion.

  • Develop new programs that:
    – Link visualization with data management.
    – Support multiresolution representations of large datasets.
    – Support simultaneous display from disparate sources.
    – Support the ability to generate and display derived quantities, and the ability to pose queries and display results.

• Develop a research program in interactive visualization with running codes that stresses the integrated design and development of coupled simulation-visualization methods.

• Establish a research program in the areas of multi-field and multidimensional data visualization.
• Automated data exploration for petascale datasets.
• Enhance life sciences visualization with particular emphasis upon the relationship with SDM.
NERSC’s Analytics Strategy

• NERSC Analytics Strategy:
  – Objective: Improve scientific productivity by increasing analytics capabilities and capacity for NERSC user community.
  – Several key strategy elements: scientific data management, visualization, analysis, support and integrated activities.

• Tactics: How to Accomplish Objectives
NERSC’s Analytics Strategy

- Broad strategic program objectives:
  - Clear picture of user needs.
  - Leverage existing and provide new visualization and analysis capabilities.
  - Enhance data management infrastructure.
  - Enhance distributed computing infrastructure.
  - Realizing analytics: support for the NERSC user community.
Leverage Existing Visualization Capabilities
Enhance Data Management Infrastructure

- **Strategic objective:** increase capability and capacity of NERSC’s data management infrastructure.

- **Tactics:**
  - Store and retrieve more bytes more quickly: global unified parallel filesystem, storage expansion.
  - Project-driven data management infrastructure:
    - Store and find data: RDBMs (record), SRM (file), FastBit (cell), others.
    - Move data: SRM (file), Logistical Networking (file), Tsunami (protocol), switched lambdas (link).
    - Share data: MDSPlus (field, variable), SRM (file).

- **Leveraging experience:** HPSS, PPDG, MDSPlus, Logistical Networking.
• Analytics is a new term and a new field.
  – It is a problem-rich research environment.
• The focus of NERSC Analytics program is not research. Program focus is on:
  – Adapting, tuning, and deploying research prototypes along with hardened technology.
  – Close interactions with CS research community figure prominently in achieving tactical objectives.
  – Working with research-grade software is often a groundbreaking activity.
• Ongoing evolution of user needs.
Conclusion

• Analytics:
  – Intersection of analysis, scientific data management, visualization, …
  – Addresses the fundamental problem of information understanding that faces modern science.

• NERSC Analytics Strategy
  – Deploy and apply constituent analytics technologies.
  – Project- and community-centric targeted impact.
Conclusion

• Objective
  – Increase scientific productivity.

\[ e = mc^2!! \]