National Energy Research Scientific Computing Center (NERSC)

Science Driven Analytics

Wes Bethel
LBNL/NERSC Visualization Group
17 May 2005

What is (Visual) Analytics?

- “A contemporary and proven approach combining the art of human intuition with the science of mathematical deduction to directly perceive patterns and derive knowledge and insight from them.”
- A methodology to confirm presence of the expected or discover the unexpected in (scientific) data.
- Martial arts analogy.

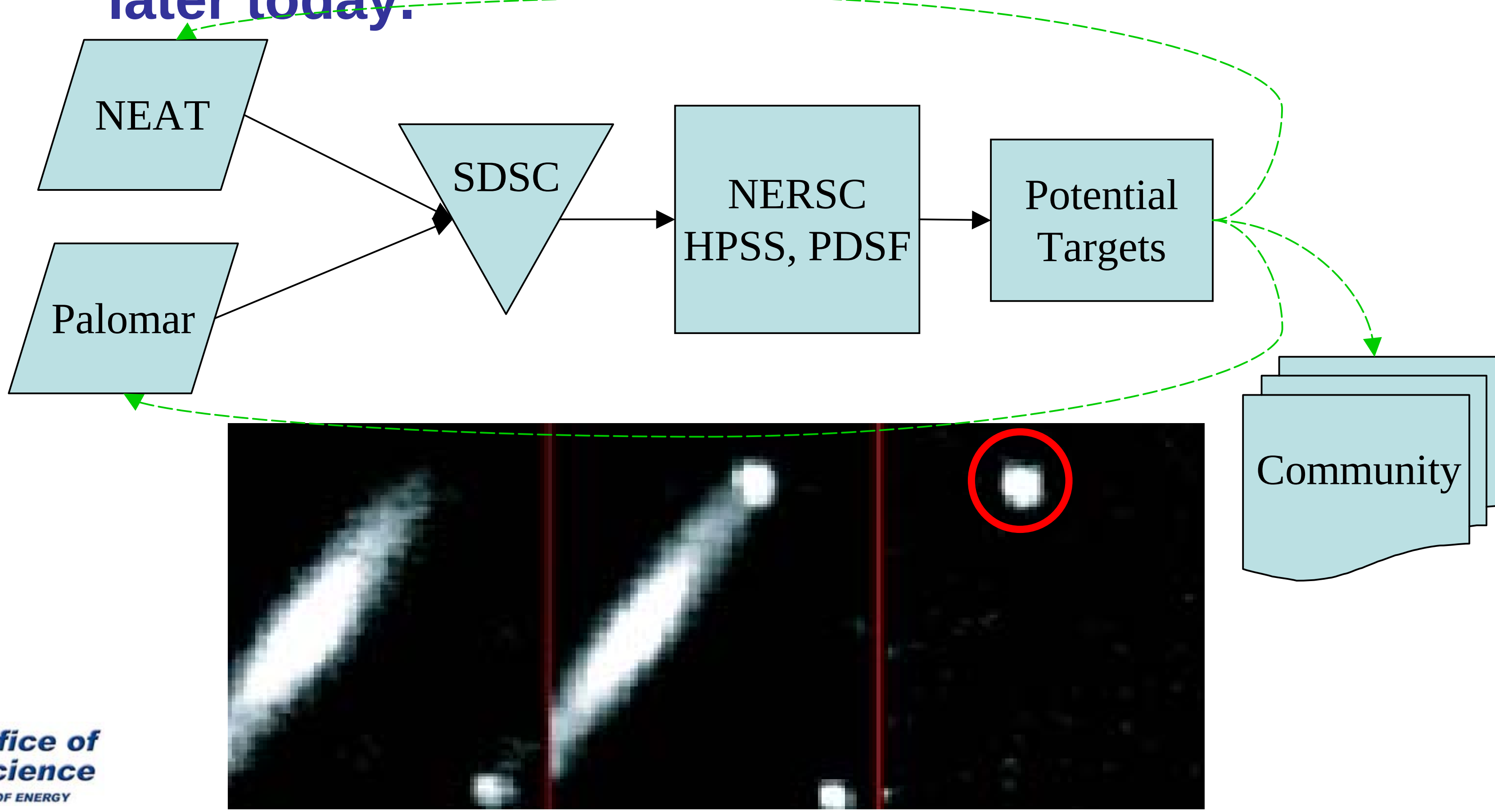
武

Why Science Driven Analytics?

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

Analytics Example

- SNFactory software pipeline
  - More information: Richard Scalzo presentation later today.



NERSC Analytics Program

From Strategic Plan:

The architectural and systems enhancements and services that integrate NERSC's powerful computational and storage resources to provide scientists with new tools to effectively manipulate, visualize and analyze the huge datasets derived from simulations and experiments.

Practically speaking:

Resources and manpower to support challenging scientific data analysis and understanding problems of the NERSC user community.

NERSC Analytics Program Elements

- **Users and their data analysis problems!**
- **Visualization**
- **Data management**
- **Workflow management**
- **Analysis**

Identify Analytics Needs

- Identify and respond to user needs.
  - Greenbook, ERCAP requests, SciDAC vis survey, Annual NERSC Survey
- Your input is crucial!
  - Please contact us directly with comments, suggestions, etc.

2002 Visualization Greenbook

Analytics Topics

- 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.

Visualization

- **Resources:**
  - New visualization/analytics platform: davinci.nersc.gov (more later).
  - Increase information visualization capabilities (software and documentation).
  - More capable and higher-throughput visualization tools (software and documentation).
  - Visualization “engineering”: remote, offline, in situ, etc.
- **Services**
  - Build on the existing visualization effort.

Data Management

- Resources
  - Facility-wide filesystem
  - Specific software packages, e.g., mySQL in the near term, others in the longer term.
- Services
  - Project-centric data formats/models.
  - Index/query capability.
  - Community-centric data caches.

Analysis

- Resources: Deploy “common analysis software”
  - Parallel R (see www.aspect-sdm.org/Parallel-R)
  - Others: PCA/ICA, clustering, discriminant analysis, correspondence analysis, spectral analysis, ….
- Service:
  - Help apply to specific projects, including integration into workflows.

Workflow Management

- Resources
  - Workflow management software (e.g., Kepler)
- Services
  - Web-brokered workflow encapsulation.
  - Assistance deploying project specific workflows and workflow components.
  - E.g., SNFactory
- Caveats
  - Emerging technology – no "one true WMS"
  - Workflow model predicated upon existence of infrastructure: data model/format, workflow component interfaces, distributed/remote component execution, data marshaling and movement, ….

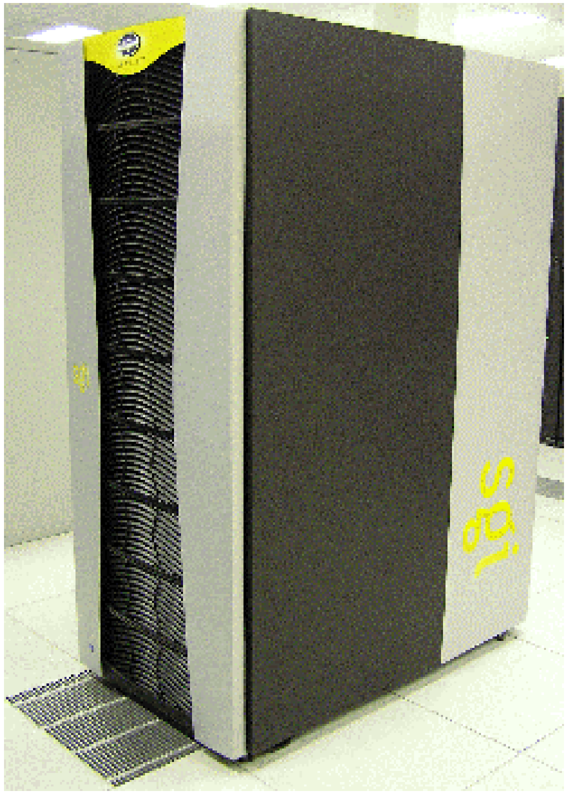
Interactive Analytics Resource: Davinci.nersc.gov

- 32P, 1.4Ghz IA64, SMP
- 192GB RAM
- 23TB scratch storage
- 10Gigabit link to internal resources (e.g., HPSS)
- 10Gigabit link to open internet.
- Coming soon: dual FC (2x2Gb/s) links to GUPFS.
- More information: http://www.nersc.gov/nusers/resources/davinci/

Davinci, ctd.

- Primary intended use: interactive visualization, analytics, analysis (data intensive applications).
  - Batch queue run at night.
- All NERSC users eligible for an account.
- Charging policy.

The End



Contact Information

- Wes Bethel, ewbethel@lbl.gov
  - NERSC Visualization/Analytics
- vis@nersc.gov
  - The entire Visualization team.
- Francesca Verdier, fverdier@lbl.gov
  - NERSC User Services
- Bill Kramer, wtkramer@lbl.gov
  - NERSC Director

Why Altix/Prism?

- Large memory SMP.
  - A serial job can access “all” memory.
- Fast internal bandwidth.
  - 6.4GB/s bidirectional NL4 interconnect fabric ideal for data intensive applications.
- Substantial high-speed scratch storage.
  - ~23TB of RAID5 disk
- Operations policy engineered to support interactive use.