STAMPEDE: A Framework for Monitoring and Troubleshooting of Large-Scale Applications on National Cyberinfrastructure

**Scientific Workflow**

- **Plan and Execute**
  - **PEGASUS**

- **Log Collection Tools**
  - `monitord` parses data in real-time
  - `nl-loader` collects workflow execution logs

- **NetLogger Tools**
  - `monitord` parses data in real-time
  - `nl-loader` stores data in a database, such as SQLite, MySQL, etc.
  - Data goes to a broker, where it can then be sent to many subscribers

- **STAMPEDE database schema**
  - Represents both the abstract workflow plan and running workflow, including the associations between the two
  - Handles parent and child workflows
  - Stores data at high granularity

**Large-Scale Workflows**
- Composed of thousands to millions of coordinated tasks
- Executed in complex distributed environments
- Difficult to track failures, search through thousands of files

**Failure Analysis Algorithms**
- Machine learning algorithms predict workflow failures based on behavior patterns
- Hard failures can be easily determined using database queries
- Soft failures are often stochastic, should be detected early for quick error recovery

**End-user Tools**
- **Stampede Python Analysis API**: Simple, uniform access to back-end database
- **Stampede-analyzer**: Quickly debug a workflow after execution is completed
- **Stampede-statistics**: Generates statistics about a running or finished workflow
  - Number of tasks/jobs/sub workflows ran/succeeded/failed/retried, ...
  - Job execution site, scheduler queuing time, execution delay, ...
- **Stampede-plots**: Interactive graphs and charts for workflow visualization

**Transform and Archive**
- **Real-time subscription**
- **Stampede DB**

**Real-time Analysis**

**Log Collection Tools**
- `monitord`
- `nl-loader`

**NetLogger Tools**
- `monitord`
- `nl-loader` stores data in a database, such as SQLite, MySQL, etc.

**PeriScope**
- Provides end-to-end system performance view to users
- Presents data from processes, hosts, and network elements using a scalable analysis and presentation framework