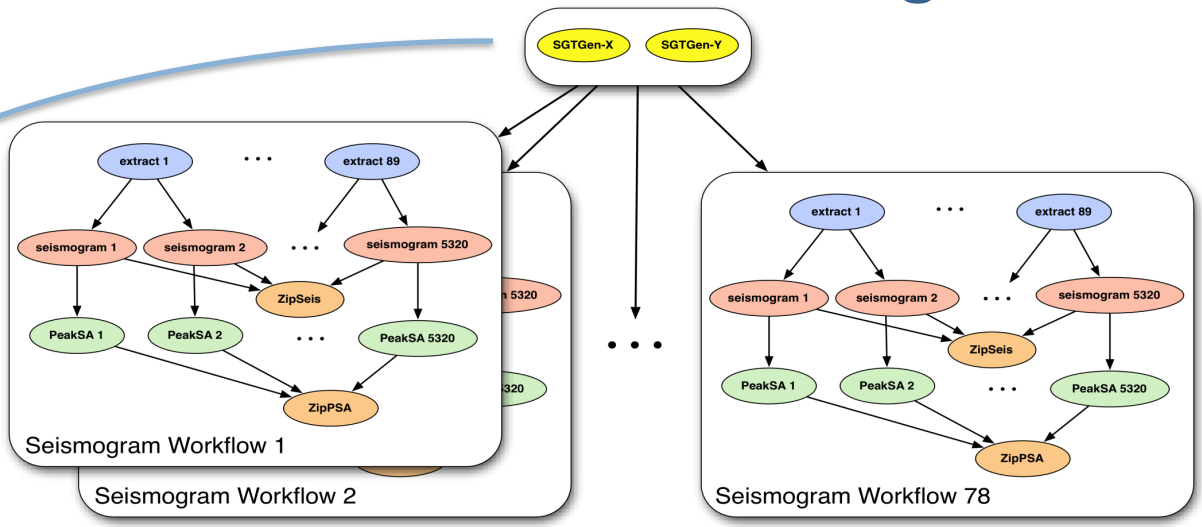
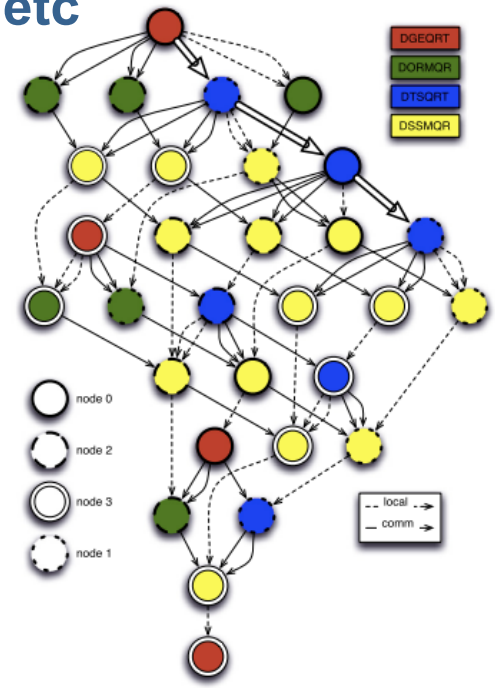
Applications and Workload Management Systems

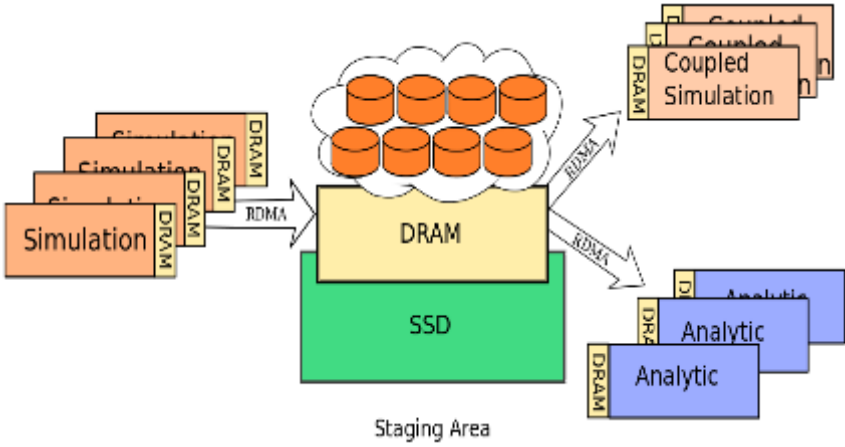


High-level workflow ensembles, SCEC Cybershake application, (Jordan et al.), managed by Pegasus

LRMS, job scheduling: PBS, LSF, etc



Data-aware Micro-task scheduling DAGuE (Dongarra et al)



In-situ coupling (Klasky et al.)

Building bridges between WMS and HPC runtime systems

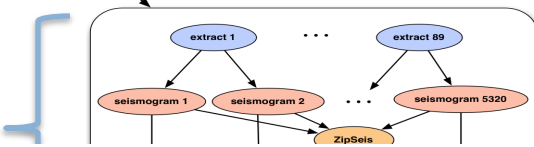
- **Task execution and Data Management: natural interfaces between WMS and the runtime systems**
  - WMS can be responsible for job scheduling and influence initial task and data placement, runtime system responsible for dynamic behavior and memory management
  - WMS can offload data no longer needed or needed on another resource for post-processing
  - Need to share knowledge between the workflow-view and the application/task view, what other jobs are coming down in the workflow? what are their data needs? what data are being produced and ready for offload?
    - Worry about energy-efficiency, picking the right systems for the jobs, making decisions about dynamic voltage and frequency scaling
    - WMS could keep track of past performance, energy usage, model future behavior,
  - Need new scheduling techniques that optimize data locality/energy consumption to make efficient use of deep memory hierarchies, and leverage high-level descriptions of the workflow structure to minimize data movement
  - Explore workflow restructuring techniques to facilitate better downstream optimizations

Possible WMS contributions to HPC systems

- **Reliability: WMS deal with: task failures, problems accessing data, resource failures, and others.**
  - Investigate how data replication techniques can be used to improve fault tolerance, while minimizing the impact of energy consumption
  - Explore tradeoffs between data re-computation and data retrieval from DRAM/ NVM/disk (time to solution and energy consumption)
- **Provenance Capture and Reproducibility: WMS capture provenance information about the creation, planning, and execution**
  - Up to now, the approach has been to save everything – problem
  - Provenance capture may need to adapt to the behavior of the application (coarse and fine levels of details, compression)
  - May want to automatically re-run parts of the computation and re-produce the results and a more detailed provenance trail on demand
- **Workflow/Applications Performance/Behavior Modeling:**
  - Understand the resource needs and behavior (performance, energy usage) of the workflow applications across scales: workflow ensemble, workflow instance, down to individual tasks and code segments
  - Need community benchmarks, execution traces and application profiles

In-roads and Futures

Single core tasks look like an MPI job



- **In-situ Pegasus**, MPI-based workflow execution: runs as an MPI program, manages sub-workflows of single core tasks
- **Resource usage characterization** of scientific workflows[1], correlations, online adaptive methods to make predictions
- **Performance modeling** of scientific workflows[2], developing analytical models for workflows (adapting HPC systems-ASPEN), integrating analytical modeling and simulation at various levels of detail, developing fault diagnostic techniques and adaptations
- **Exploring the role of WMS lookahead** when provisioning resources ahead of the workflow execution[3]
- Need to **figure out the interactions** between the various layers of workload and data management software
- Need to consider **wide area network management** (SDNs) and Science DMZ/Data Transfer nodes as part of the overall system
- **Build collaborations** between researchers with varied expertise
- **Should not forget reality on the ground:**
  - HPC/workflow programming is still a craft and hard to do (**Debugging!**),
  - Systems today have policies that make them hard to use (firewalls, limited remote jobs submission interfaces, very limited max walltimes)
  - Many smaller-scale applications that can benefit from new solutions

[1]dV/dT: Accelerating the Rate of Progress Towards Extreme Scale Collaborative Science (DOE)
[2]Panorama: Predictive Modeling and Diagnostic Monitoring of Extreme Science Workflows (DOE)
[3] ADAMANT: Adaptive, Data-Aware, Multi-Domain Application Network Topologies (NSF)