Finding Tropical Cyclones on a Cloud Cluster: Using Parallel Virtualization for Large-Scale Climate Simulation Analysis

Daren Hasenkamp*

Alex Sim, Michael Wehner, Kesheng Wu

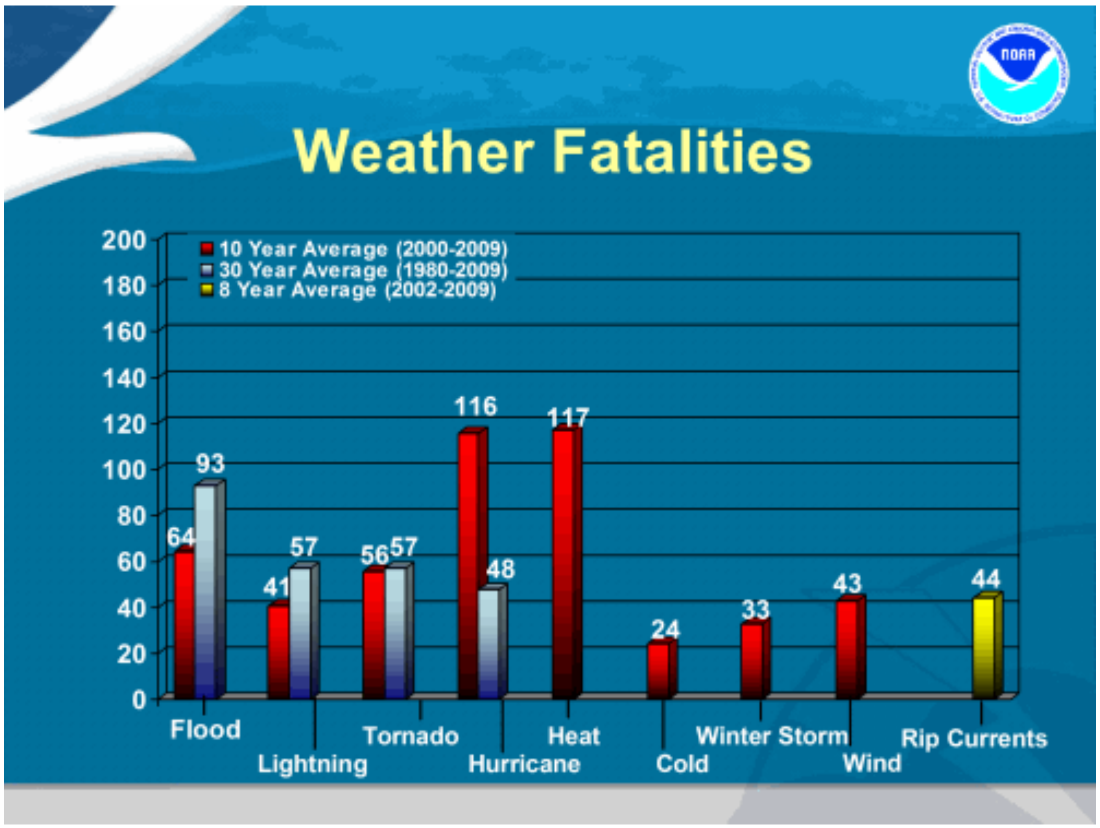
Lawrence Berkeley National Laboratory

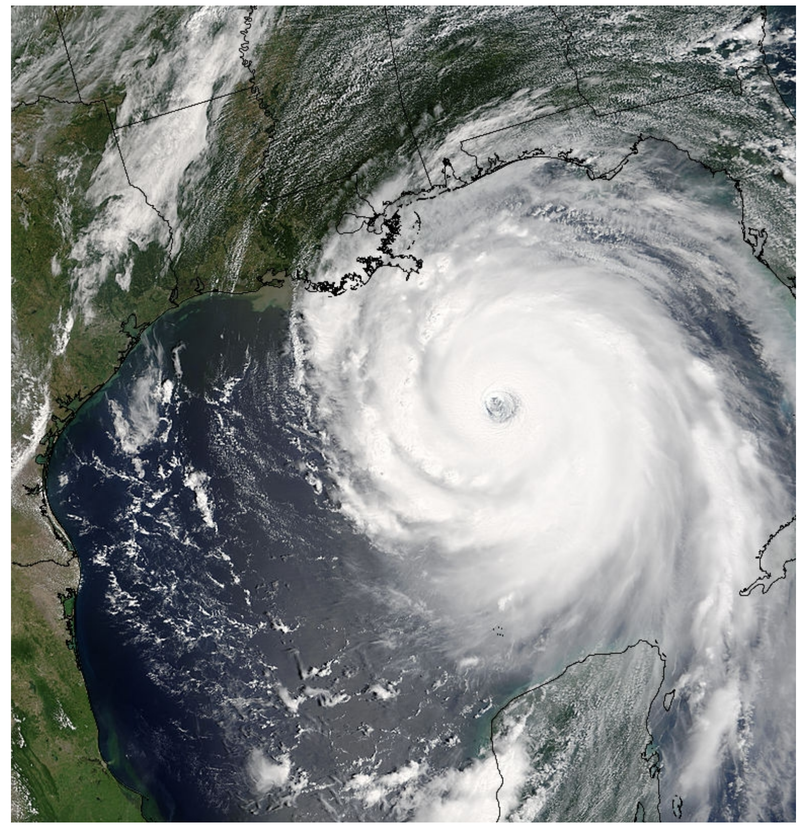
*University of California, Berkeley

Why Study Tropical Cyclones?

Tropical cyclones are among the most deadly natural phenomenon

Climate change could increase the frequency of severe tropical storms





[Weather fatalities from weather.gov]

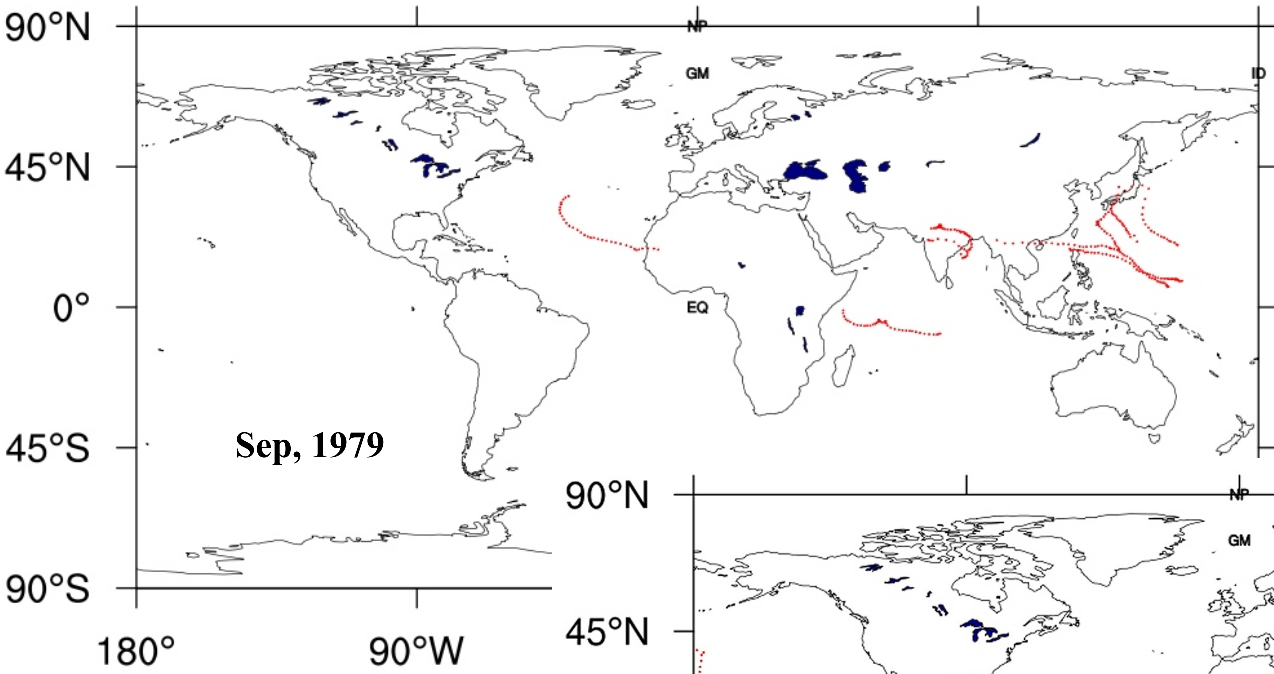
Predicting Tropical Cyclone Statistics

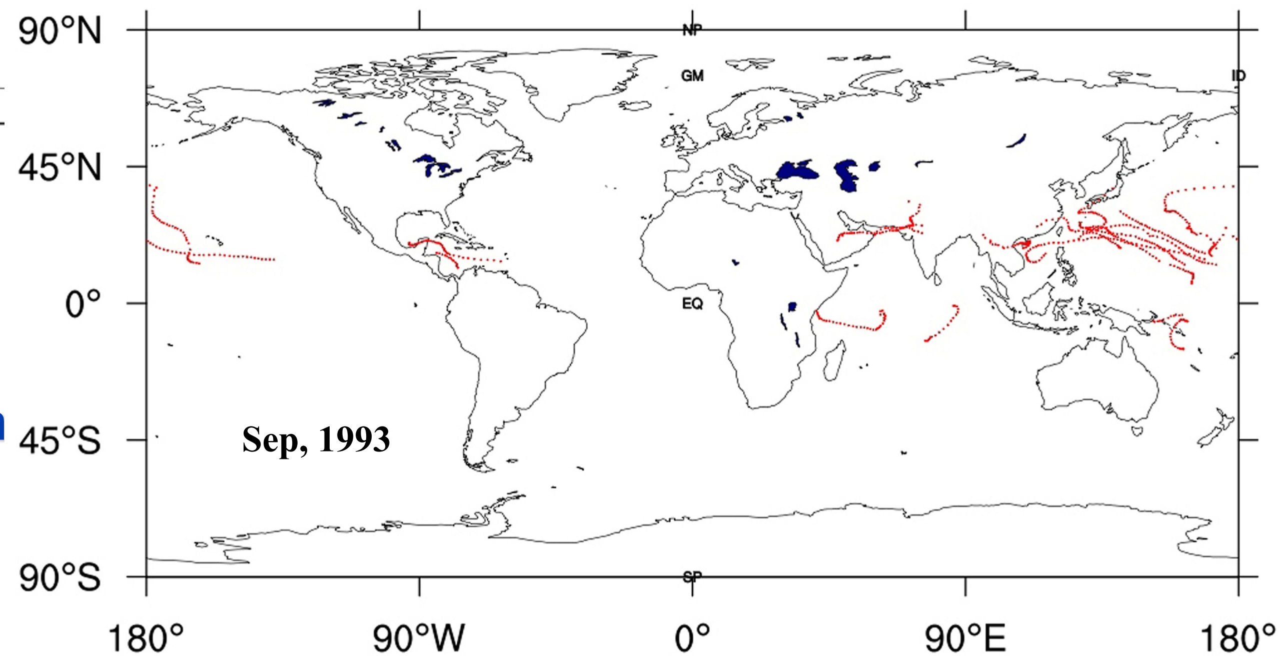
- Climatological study: Predicting statistics of tropical cyclones, not any individual storm
- Approach: simulate climate in the future, gather statistics from simulation data
- Case study: fvCAM (finite volume version of the Community Atmospheric Model) dataset (version 2.2)
  - 15 simulated years with 6 hour time steps
  - Mesh point resolution of 0.5 degree latitude by 0.625 degree longitude
  - Roughly 500 GB, 1000 netCDF files
  - Scientists will run this simulation for 100 simulated years with many different initial conditions, generating many terabytes of raw data

TSTORM code

- **TSTORM code used to track tropical storms**
  - **Based on the criteria established by Knutson, et al. from Geophysical Fluid Dynamical Library (GFDL), 2007 BAMS 88:10 1549-65**
  - **Searches for high vorticity, local pressure drop, and warm core**
    - **A local relative vorticity maximum at 850 hPa exceeds $1.6*10^{-4}$ s^{-1}. Vorticity is the curl of wind velocity, and s is time in seconds.**
    - **The surface pressure increases by at least 4 hPa from the storm center within a radius of 5 degrees. The closest local minimum in sea level pressure, within a distance of 2 degrees latitude or longitude from the vorticity maximum, is defined as the center of the storm.**
    - **The distance of the warm-core center from the storm center does not exceed 2 degrees. The temperature decreases by at least 0.8 degrees Celsius in all directions from the warm-core center within a distance of 5 degrees. The closest local maximum in temperature averaged between 300 and 500 hPa is defined as the center of the warm core.**

Tropical storms





- **Produced by TSTORMS using virtual machines on cloud computing facility**

TSTORMS code and Parallelization

- **TSTORMS**
  - A single thread sequential program
  - Running on a single processor
  - Analysis of 500GB of simulation output can take several days
  - Need to analyze many petabytes, but can not wait for decades
- **Parallelization is needed**
  - Running multiple TSTORMS processes, one for each time step
- **Challenges in traditional parallel processing**
  - Need to rewrite the code with MPI
  - Port dependent software libraries and run-time systems
- **Cloud computing as an alternative**
  - Using virtual machines to package existing analysis code, libraries and run-time systems, no need to rewrite code
  - Portable to many computing hardware

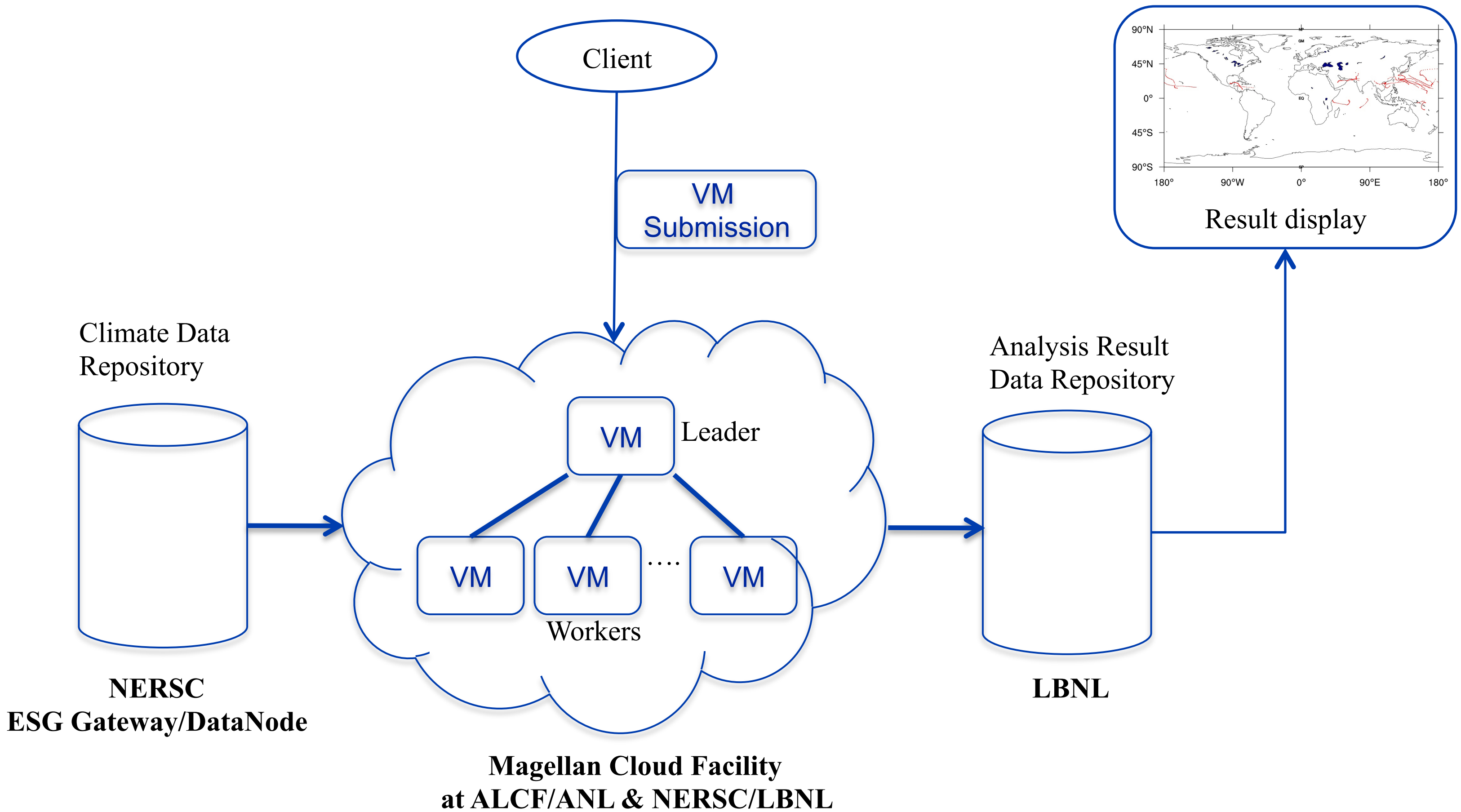
Three Different Approaches

- Virtual machine on cloud computing
  - Eucalyptus VM submission
- Virtual machine on grid computing
  - Pre-loaded VMware image
- MPI parallel processing on cluster computing
  - Needed code re-write for MPI and local compilation

Virtual Machine Coordination

- **Difficulties in controlling virtual machines instance**
  - Hard to control exactly how many virtual machines instances are launched. For example, a user requesting 40 instances might only receive 36. Not all cloud clusters share this property, but it was our experience during the tests.
  - Virtual machine instances launch at varying times: If a user makes a request for 20 VM instances, the first instance might start a half hour before the final.
- **MPI-based process coordination for data-driven parallelism comes easier.**
- **Needs of VM analysis coordination**
  - Coordination through leader election
  - Coordination through external service

Analysis with virtual machines on cloud computing



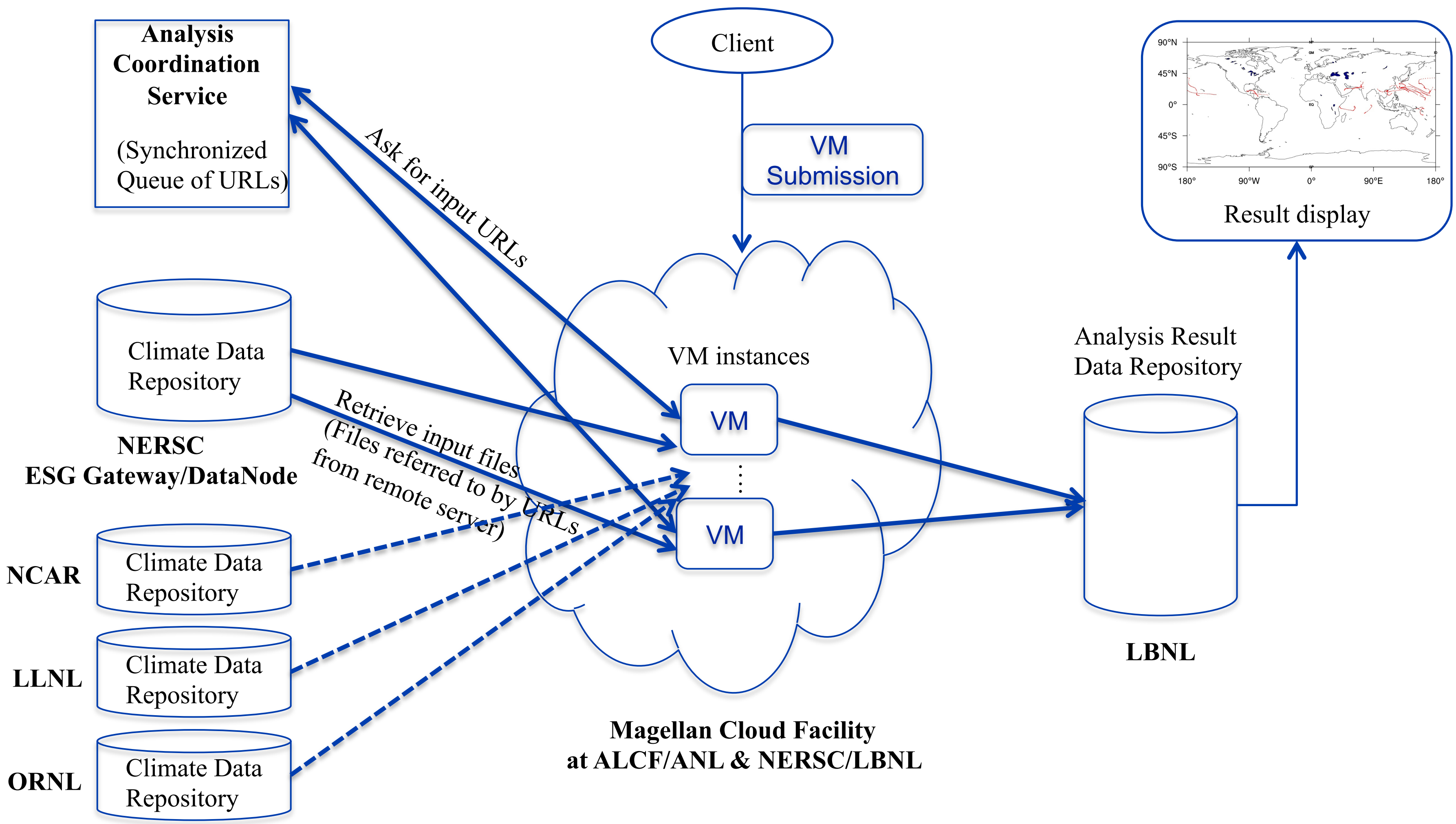
Coordination using Distributed Leader Election

- **Leader election**
  - elect one VM instance as a leader at launch time
  - track job status and coordinate VM instances
  - leader maintains a synchronized queue of URLs to input files from which all other VM instances pull one URL at a time.
  - Advantage: the job is self-contained
    - A user can launch many instances, and does not have to perform any further tasks, such as setting up a remote service.
  - Disadvantage:
    - Static input URLs
    - Difficulties in dynamic coordination for multiple source repositories
    - Dependency on the leader instance on the particular node

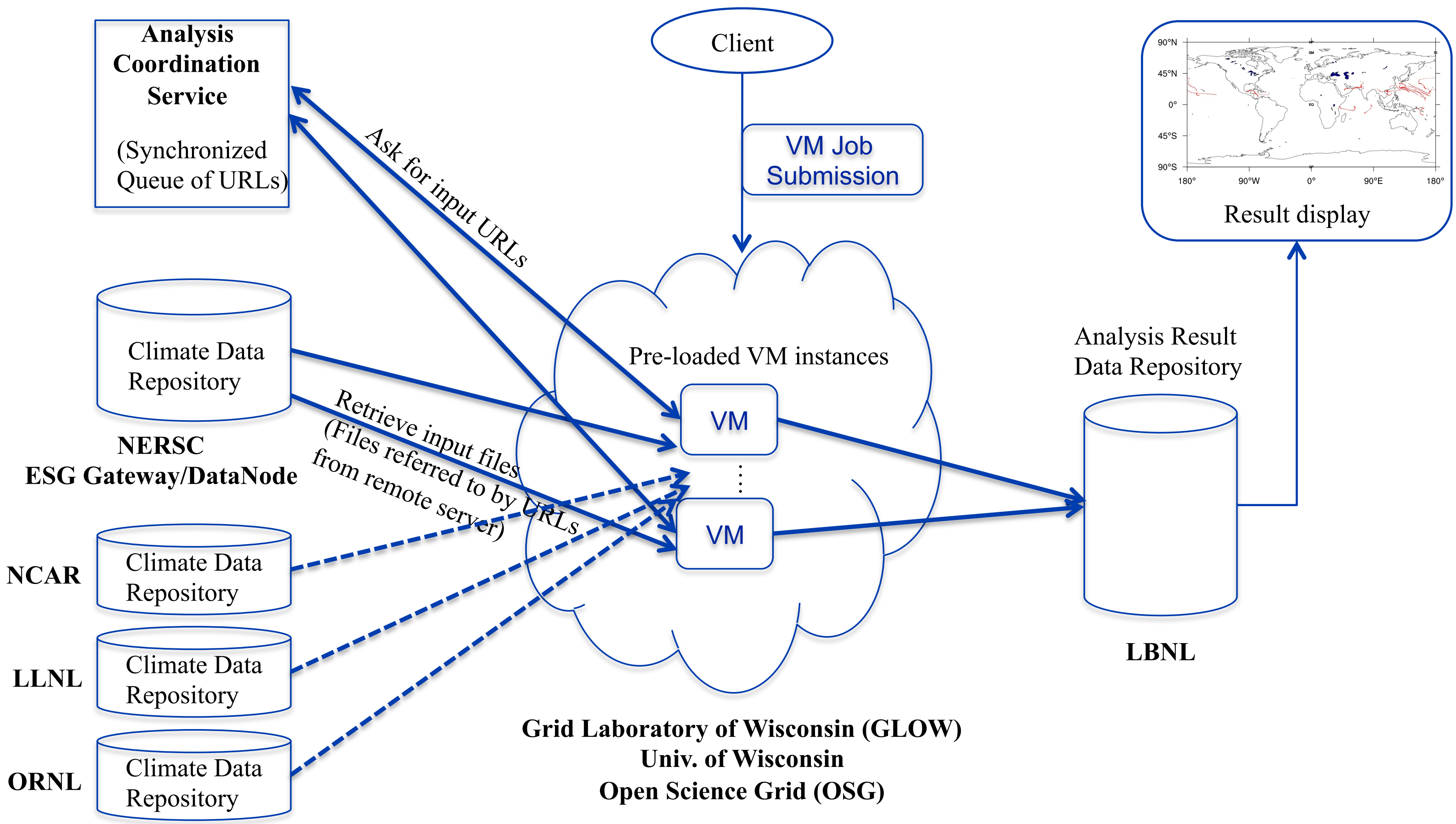
Coordination through a Remote Service

- External analysis coordination service
  - Service maintains a synchronized queue of URLs to input files from which all other VM instances pull one URL at a time.
  - Advantage:
    - Easy setup
    - Dynamic coordination for multiple source repositories
  - Disadvantage:
    - Dependency on the remove service

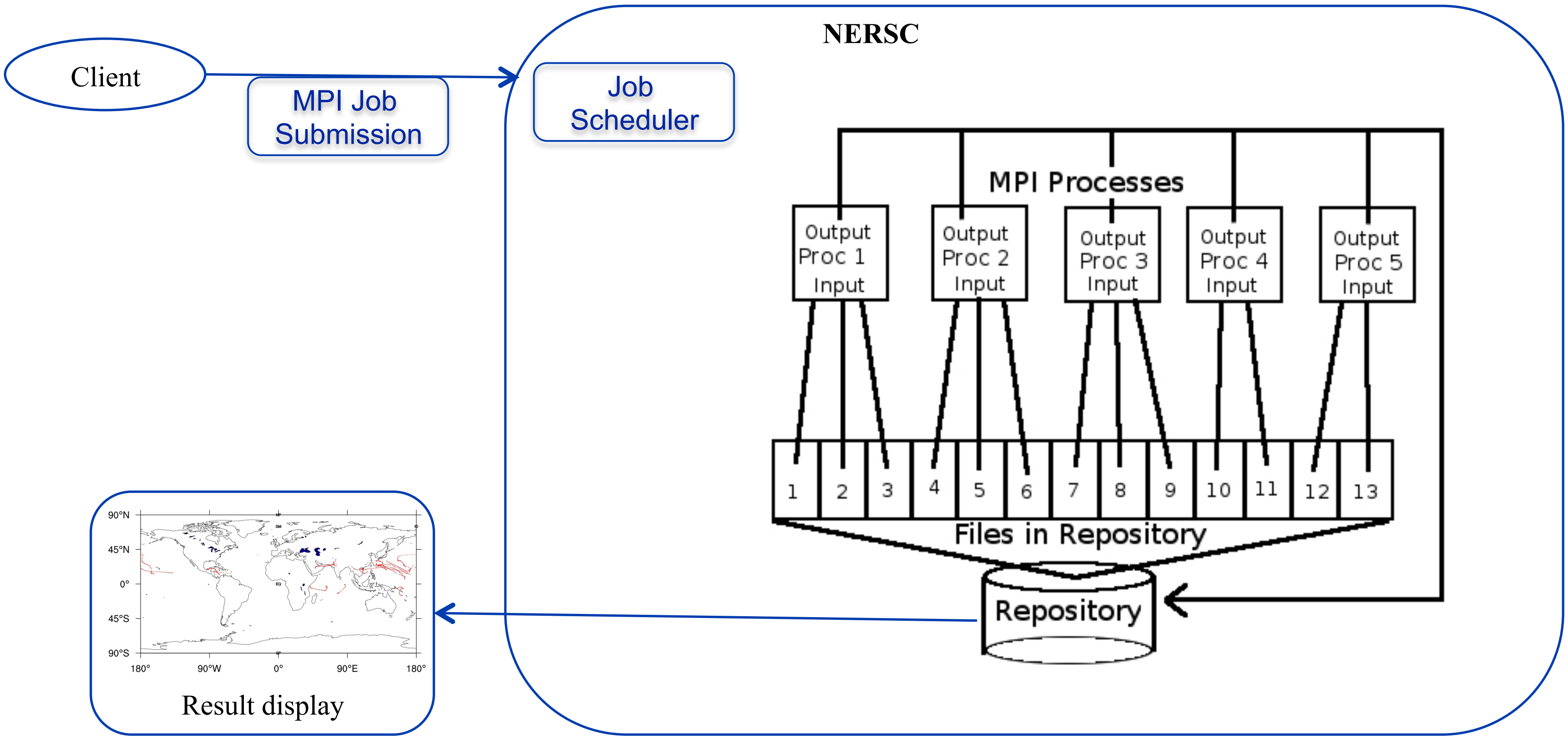
Analysis with Virtual Machines on cloud computing



Analysis with Virtual Machines on Grid computing



Analysis with MPI parallel processing on Clusters



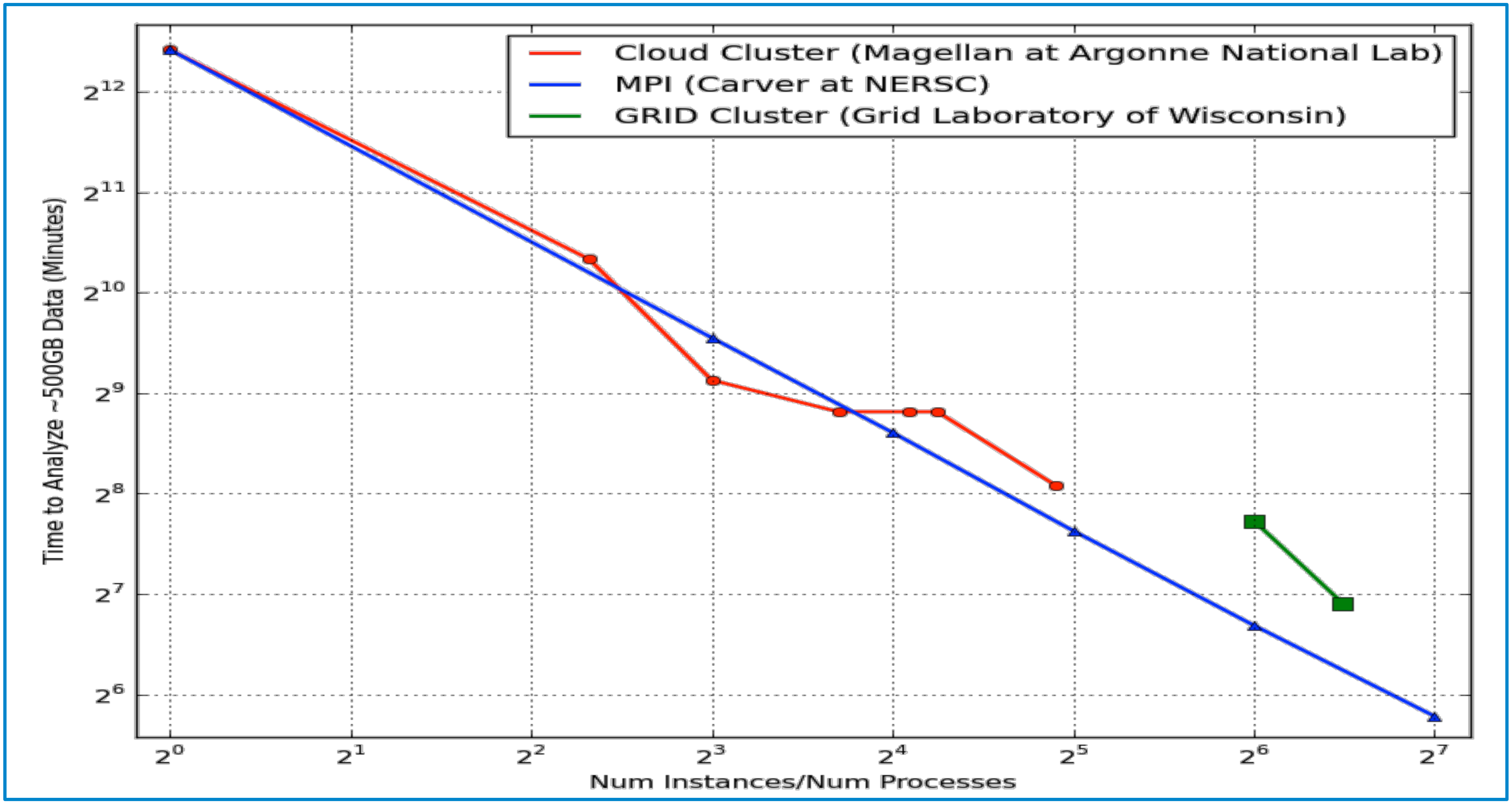
Test setup

- **Magellan cloud and Carver cluster**
  - each node on each system contains dual quad-core Intel Nehalem 2.66GHz processors and 24GB RAM
- **GLOW**
  - GLOW nodes we used utilized Xeon 2.66GHz and 3.2GHz processors, and had enough RAM for TSTORMS to execute without using virtual memory
  - Our VM on GLOW had compute resources comparable to, though not exactly the same as, instances on Magellan and processes on Carver.
- **Source data on GPFS at NERSC**
  - Runs on Carver had somewhat of a speed advantage over VMs since data could be accessed through a local file system rather than needing to be sent across a network.
  - Disadvantage from virtualization overhead on VMs compared to Carver MPI processes.

Results (1)

- **Performance from VM-based analysis comparable to MPI-based analysis**
- **In one test, Magellan VM-based analysis actually performed better than Carver MPI-based analysis**
  - **Analyzing our 500GB repository on Carver using 8 processes took 3 hours longer than on Magellan using 8 virtual machine instances (~12.5 vs ~9.5 hours)**
- **Using 30 VMs, analysis of the 500GB dataset in ~4.5 hours**
  - **Using a workstation with similar computational power, it can take several days; roughly 100 hours**
- **Analysis in ~2 hours using 90 instances on GLOW**
  - **Conveniently short amount of time for a scientist to wait for analysis output, and it is comparable to analysis speed on Carver**

Time v. Number of Processes



Results (2)

- Total analysis time as a function of number of instance or number of processes
  - On Carver,
    2 * (the amount of processes) → ½ (total analysis time)
  - Using VMs on a cloud, this holds only approximately
    - Expected that VM instances can have different starting times, whereas processes in MPI start almost at the same time
    - Effects of shared network
      - Our VM runs somewhat faster late at night and on weekends, when there is less traffic on network resources.
      - The anomalous 8-instance test on Magellan was started on a Friday night, and competition for both network bandwidth and cloud nodes would have been relatively low.

Conclusion

- **Test analysis took 5-7 days on a workstation to ~3 hours on 32 VMs on Cloud**
- **Analysis performance on cloud computing is comparable to analysis performance on MPI-based batch computing**
  - **MPI jobs are more predictable in performance**
  - **Variability on Cloud jobs is larger**
    - **Successful number of VM initialization varies**
    - **Network performance for remote data access**
    - **Storage capacity and performance**
- **Parallel virtualization**
  - **A viable paradigm for large-scale data analysis**
  - **Offers an attractive environment**
    - **analysis programs can be configured once and run anywhere with configurable, and potentially massive, levels of parallelism and efficiency, comparable to a traditional batch-based computing system**