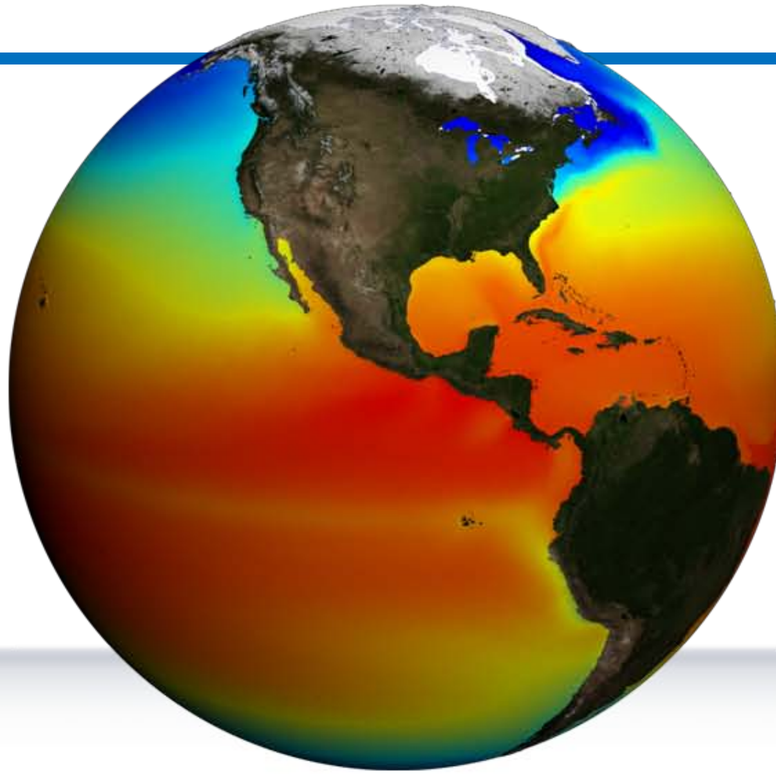
HPC at NOAA/GFDL

Bernie Siebers
HPC Lead at GFDL
May 6, 2013

NOAA's Mission



To understand and predict changes in Earth's environment and conserve and manage coastal and marine resources to meet our nation's economic, social, and environmental needs

NOAA Is Vital to American Economy

A third of the GDP ($4 trillion) is reliant on accurate weather and climate information.

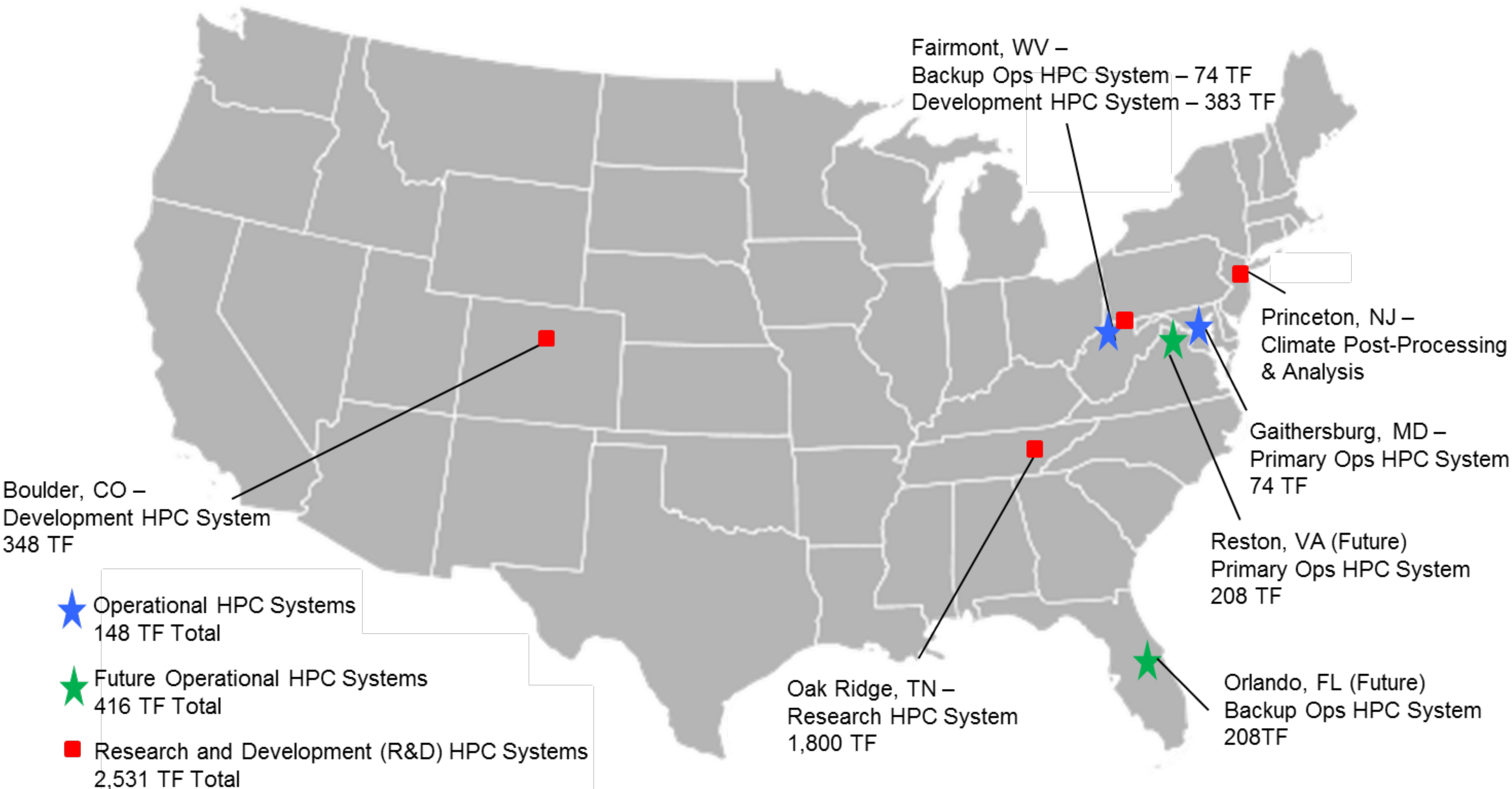
Example of NOAA's role:

NOAA provides weather, water, and climate forecasts and warnings for the private and public sectors. Annually, NOAA provides 76 billion observations, 1.5 million forecasts, and 50,000 warnings.

-NOAA provides economic benefits of $240 million per year in mitigating flood losses
-NOAA's aviation forecasts reduce aviation delays and save the industry $580 million per year.

NOAA High Performance Computing

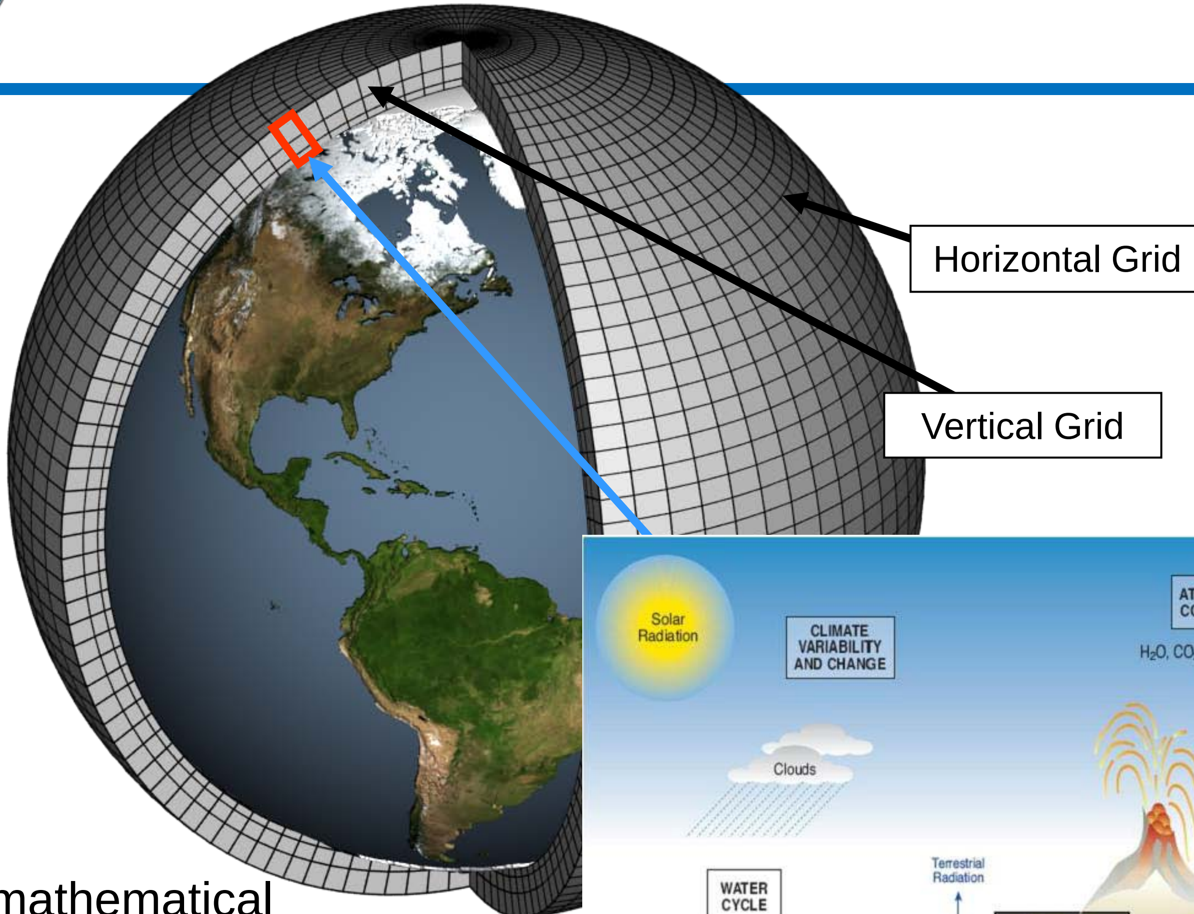
Capabilities & Locations

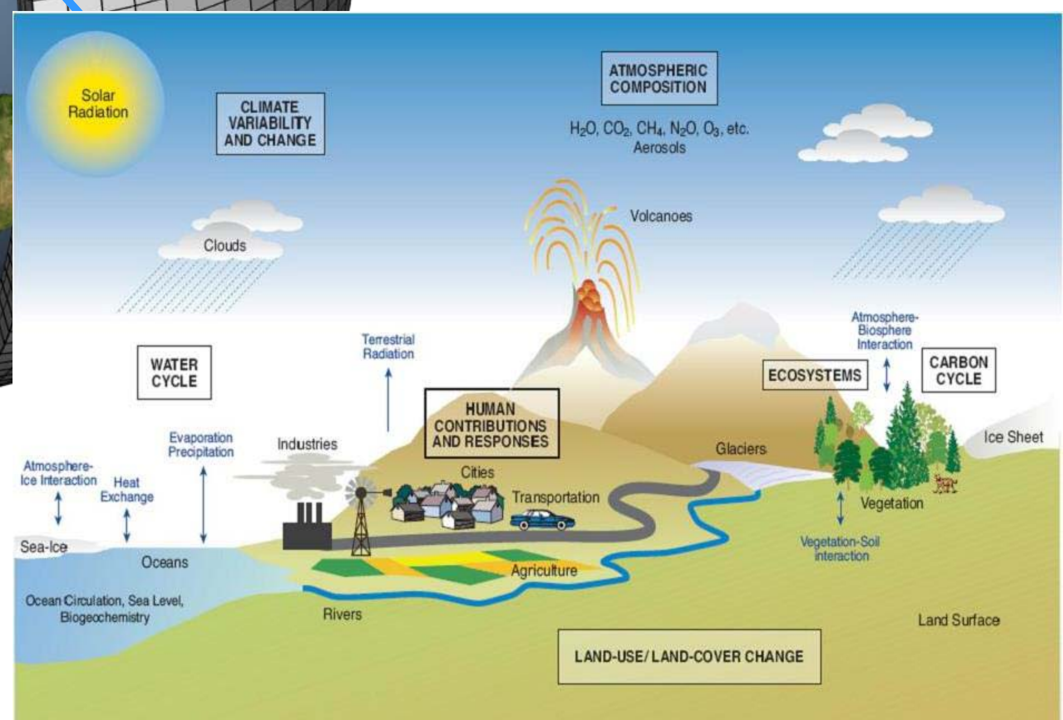


GFDL Climate Computing

- GFDL's primary focus is coupled atmosphere-ocean climate simulations, many running for hundreds of model years
  - GFDL: Geophysical Fluid Dynamics Laboratory, Princeton, NJ
  - GFDL participates in IPCC, Intergovernmental Panel on Climate Change
- Models run on NOAA's Gaea and Zeus, and other machines
  - Gaea: 120K core 1.8PF Cray XE6 at Oak Ridge, Tennessee
  - Zeus: 28K core 384 TF SGI ICE at Fairmont, W.Virginia
- Model data processed and stored at GFDL
  - Transmitted to GFDL via two 10Gbit NOAA WAN links
  - Processed into time series and time averages on 800 core Dell cluster
  - 42 PB long term data stored on 35000 Oracle T10K-B and T10K-C tapes in 4 SL8500 silos, managed as an HSM by SGI's DMF

Schematic Global Climate Model (GCM)

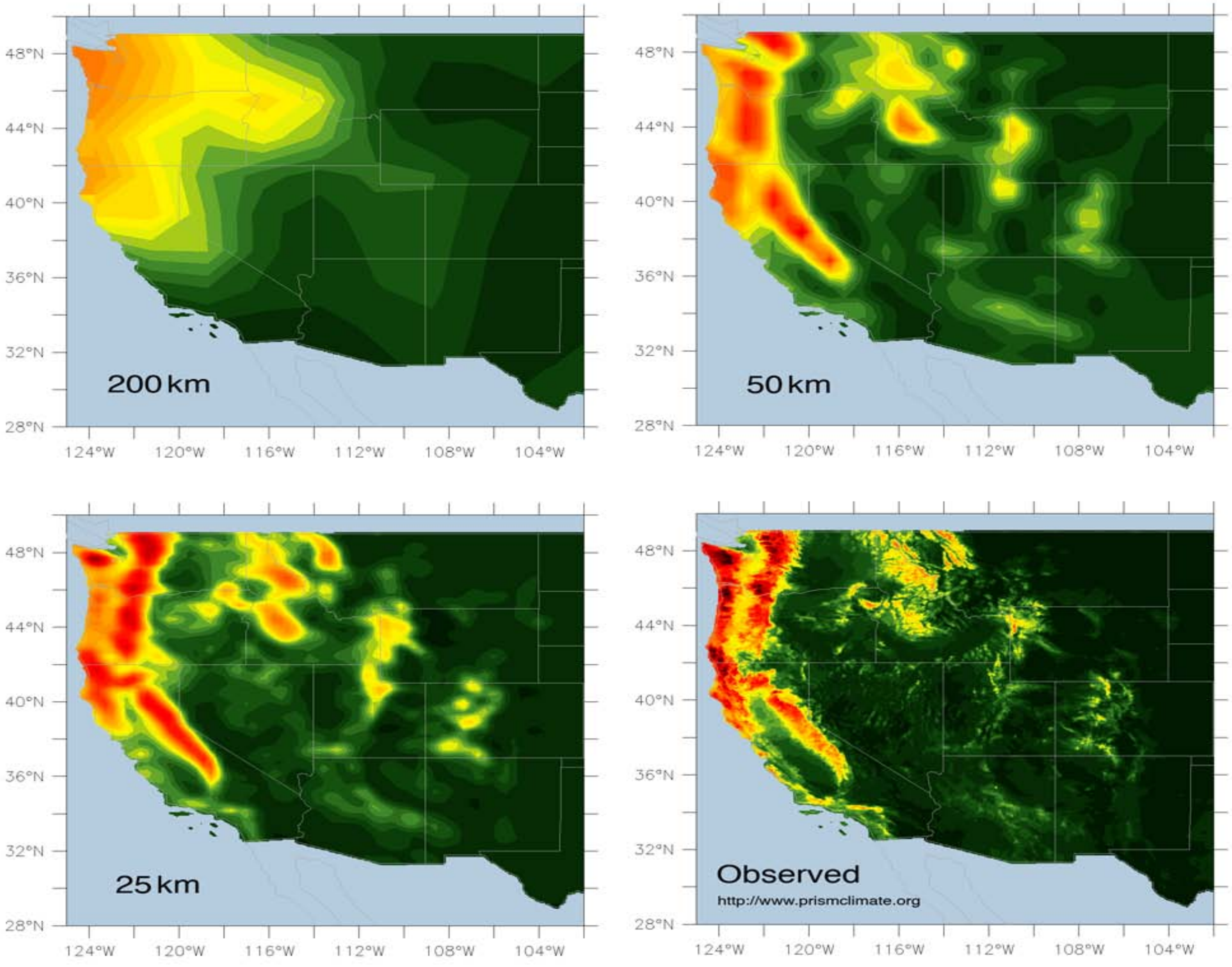




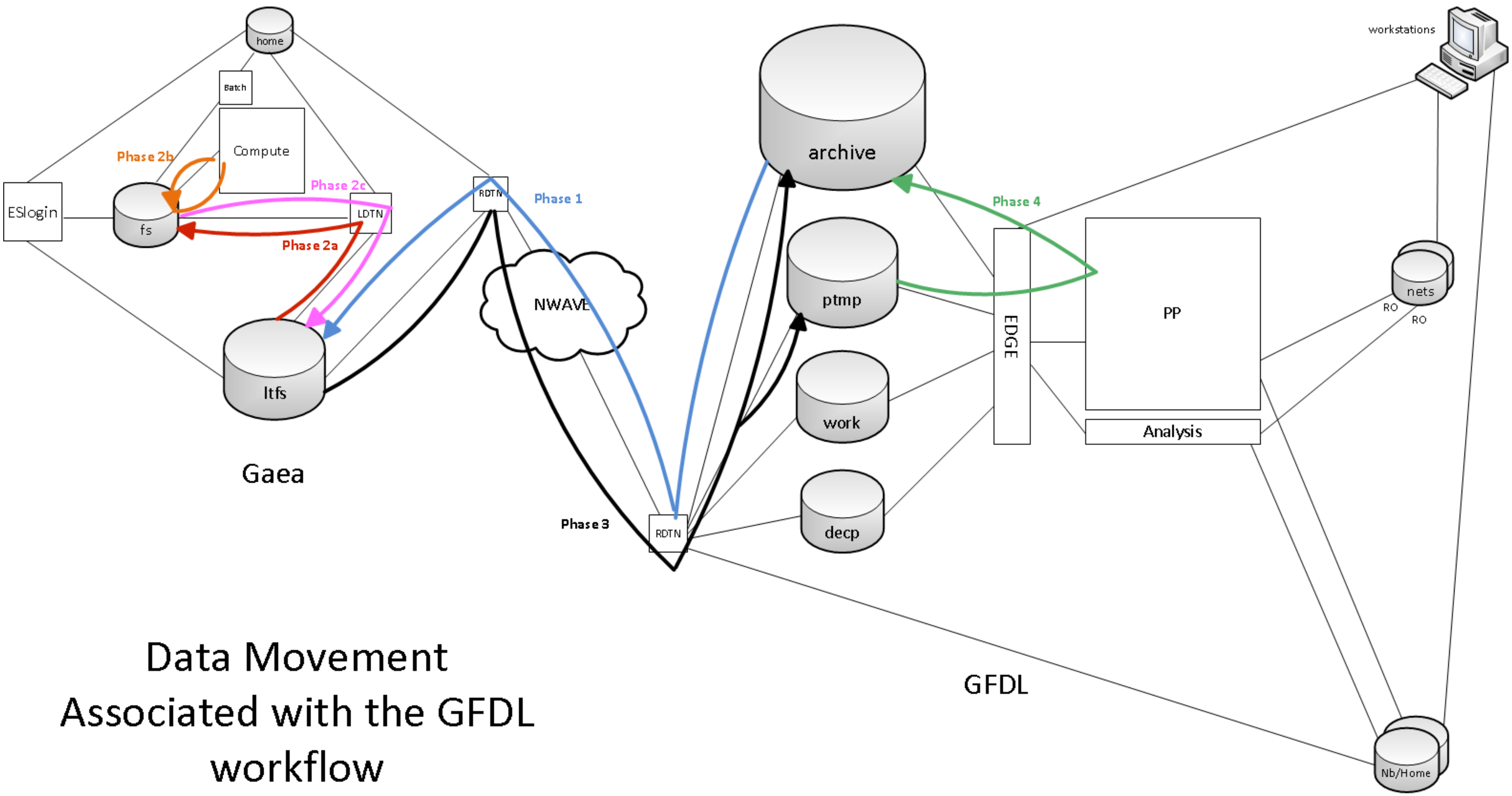
A GCM is a mathematical representation of the major climate system components and their interactions. The GCM equations operate on a global grid and are solved on a computer.

Data Volume

Model Grid Size



Workflow



Data Movement
Associated with the GFDL workflow

Long Runs, Big Data

- GFDL has been doing long runs with big data for decades
  - GFDL has done ocean and atmosphere modeling for over 50 years
  - A climate experiment may run for 6-12 months, 200+ model years
  - User checkpoints break experiments into ~16 hour segments
  - Any error affecting a restart snapshot alters the rest of the run
  - To guard against such problems, GFDL scientists watch closely for any data reproducibility or integrity issues
- This caution arose out of hard experience
  - In the 1970s, GFDL scientists lost over a year of productivity when one of the first vector computers, with early printed circuit boards, kept flipping bits and had to be rebuilt on site
  - **Nearly every machine since has experienced some data integrity issue**

A Few Historical Examples

- CPU design problems
  - Microcode timing error: failed to wait for prior instruction completion
  - Hardware instruction error: given one instruction sequence, with just the wrong timing, failed to wait for a memory fetch to complete
  - Microprocessor error: flipped low-order bit under certain voltage and temperature conditions, well within normal operating range

- Broken hardware
  - I/O processor dropped one 16-bit packet, shifted rest of stream
  - Disk controller fetched data from wrong disk address
  - Disk controller in error recovery occasionally wrote its firmware to disk

A Few Historical Examples

- Software errors
  - Compiler failed to insert “wait for sync-to-memory” instruction, so that occasionally one would fetch un-updated data on next instruction
  - Software copy package we obtained from another lab would occasionally shuffle the order of blocks
  - Flushing to disk or tape supposed to have completed, but has not
  - Application code failed to set or check semaphore

- Note some frequent characteristics
  - Low-order bit (high order bits more likely already detected and fixed)
  - Use of wrong memory location or data, but nearby or just not updated
  - Often highly sensitive to timing or instruction sequence
  - Consequence: errors are infrequent, subtle

How GFDL Protects Itself

- Dual runs
  - Dual run most jobs on new machines for 30 days, and after that for a small fraction of jobs, and jobs critical to get exactly right
  - Scientists take care to provide for a mode, even on a parallel processor, where dual runs bit reproduce

- Checksums
  - End-to-end checksums on WAN transfers, PLUS packet checksums
    - Users have an option to turn off end-to-end checksums
  - Tape data: insist on perfect write, get two bits for error correction
  - Enable application checksums where possible, including HSM

- <u>Bottom line: don't trust – verify</u>

Thank You