HPC Performance Advances for Existing US Navy NWP Systems

Timothy Whitcomb, Kevin Viner
Naval Research Laboratory Marine Meteorology Division
Monterey, CA

Matthew Turner
DeVine Consulting, Monterey, CA

ECMWF HPC Workshop 2016

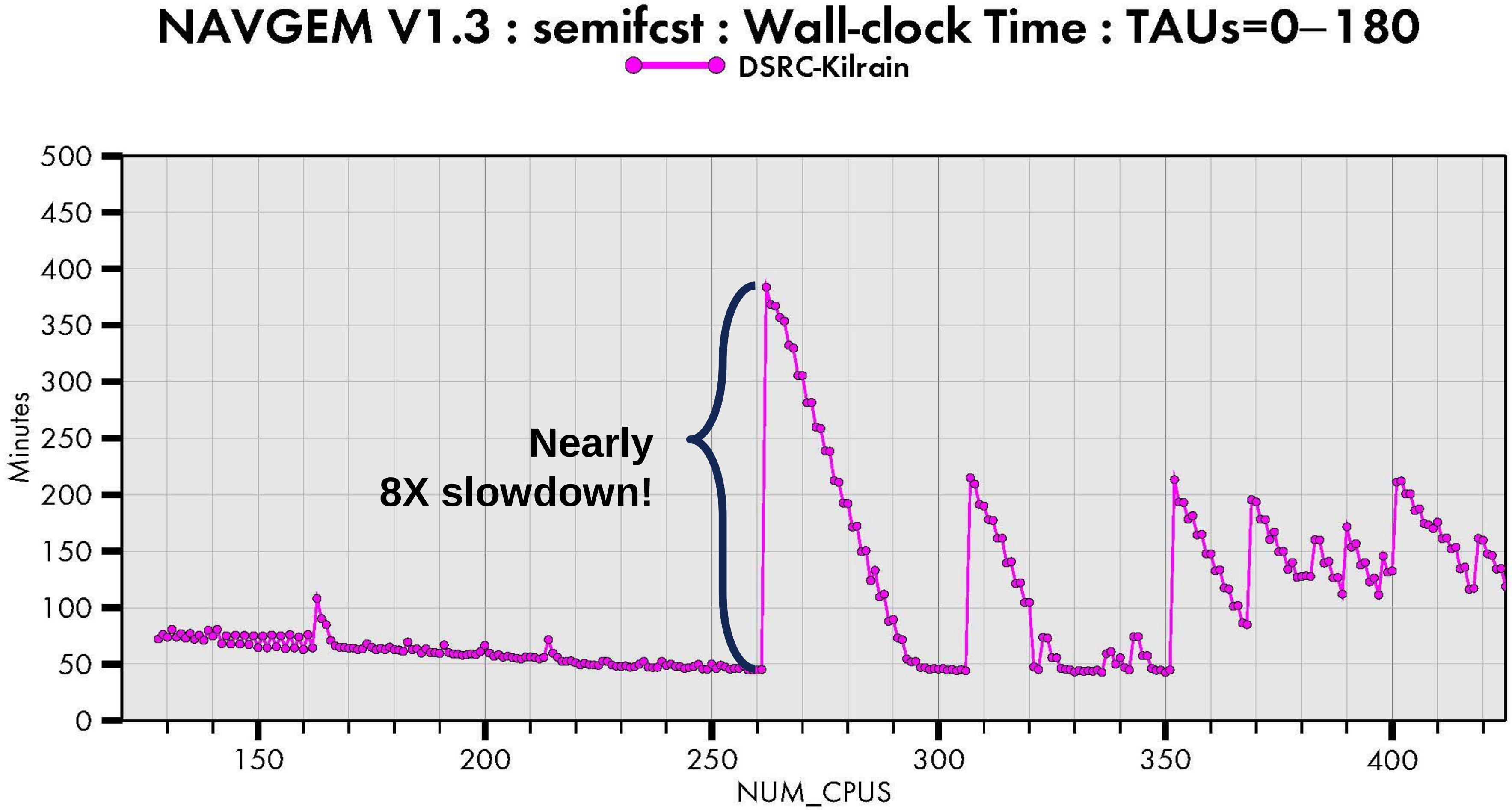
Distribution Statement A: Approved for public release; distribution is unlimited

Significant Upgrades to Navy NWP

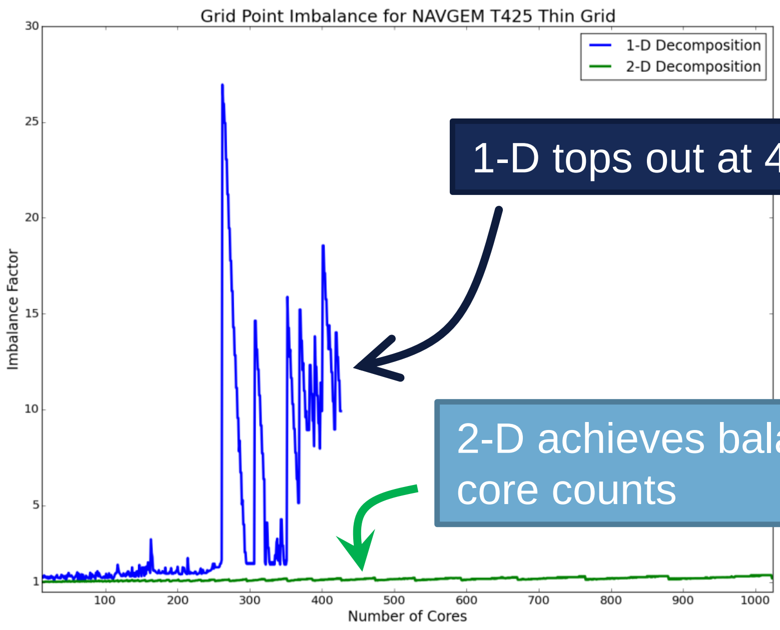
- Focused on current global system
- Allow for higher resolution and prepare for next-generation system
- Three focus areas
  - New 2-D MPI domain decomposition
  - Asynchronous I/O (implemented via ESMF)
  - Improved end-to-end system performance via task-based parallelism and workflow managers

U.S. NAVAL RESEARCH LABORATORY

New model infrastructure updates

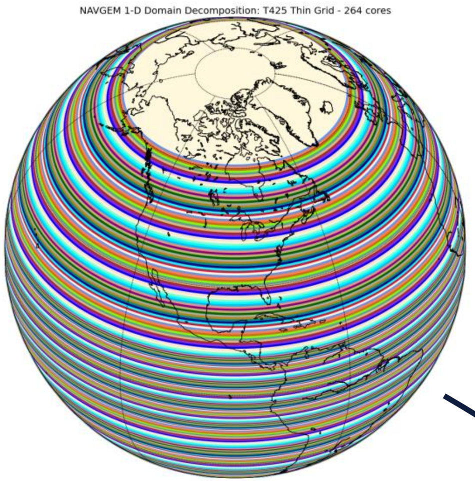


Future scalability challenges identified with thinned Gaussian grid

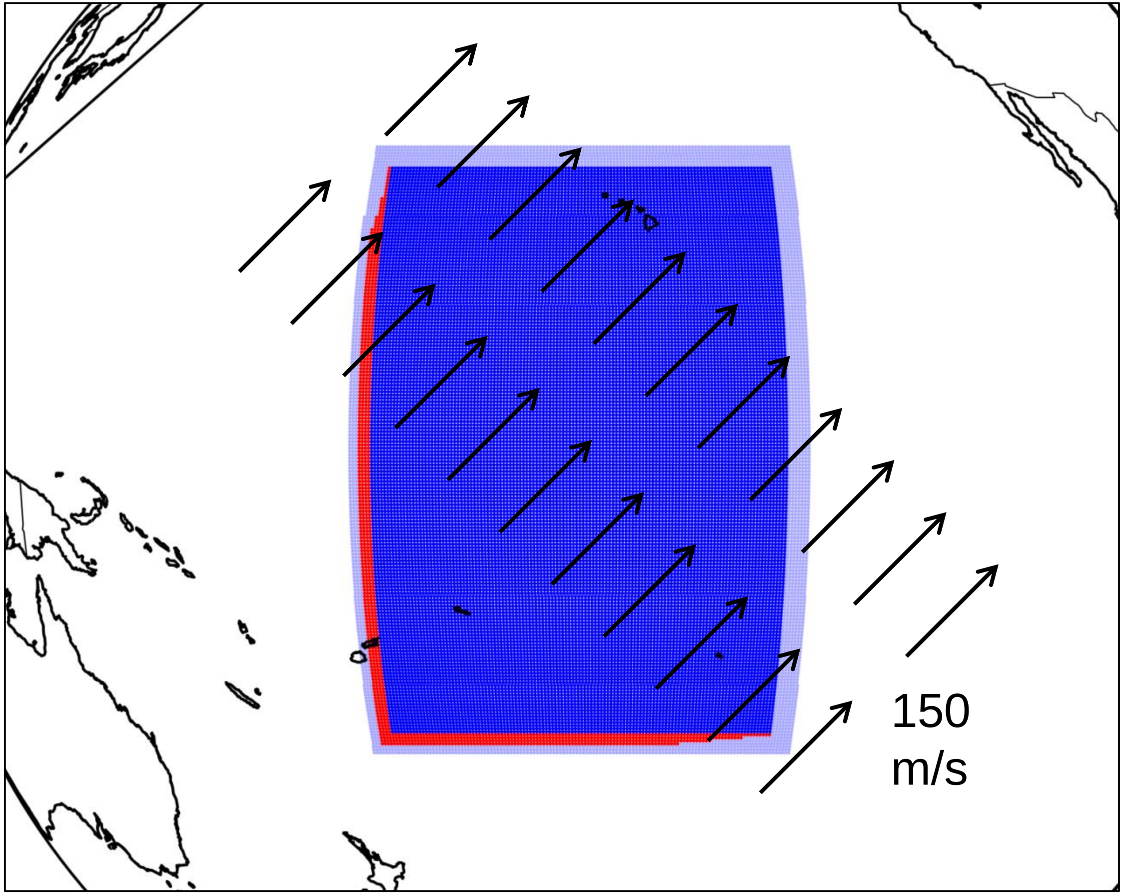


1-D tops out at 426 cores (for T425)

2-D achieves balance at higher core counts



Current 1-D domain decomposition shows significant challenges to grid point load balancing after 264 cores which is almost entirely mitigated with the 2-D decomposition.



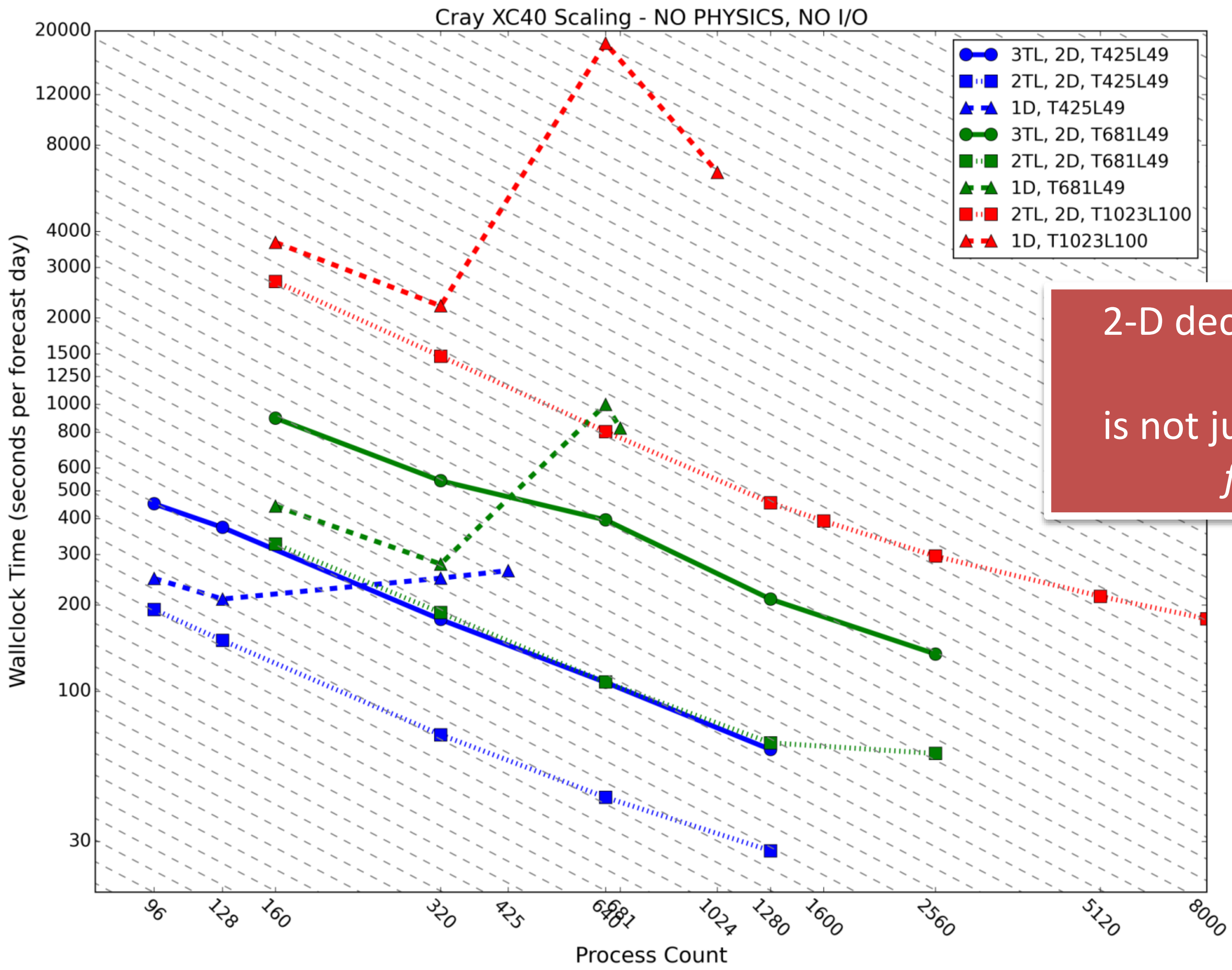
Semi-Lagrangian halo exchange was modified to use MPI-2 RMA operations.

- T425
- 24 Cores
- Uniform wind field
- **Light Blue**= maximum halo
- **Red** = communicated points
- **Blue** = on-process points

Initial Optimization

Optimizations	Runtime Reduction
Transform MPI Derived Datatype	7% - 37% of overall (varies by core count)
Legendre Transform Optimization	~20% of transform runtime (~4-5% overall)
Halo Generation	20-36% of semilag (~10-18% overall)
Vectorization / Cleanup of legacy code	~12% of overall
Reduced Cubic Interpolation (Poorman)	~19% of overall (compared to full cubic)
MPI Halos	~10% overall (compared to coarray)

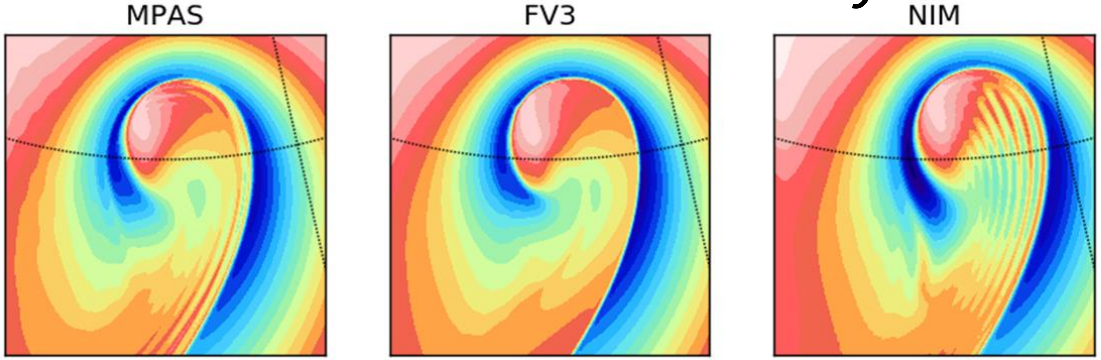
Additionally, update MPI to remove many broadcasts (especially in initialization), create windows once at startup, and replace single send/recv call with non-blocking send & receive.



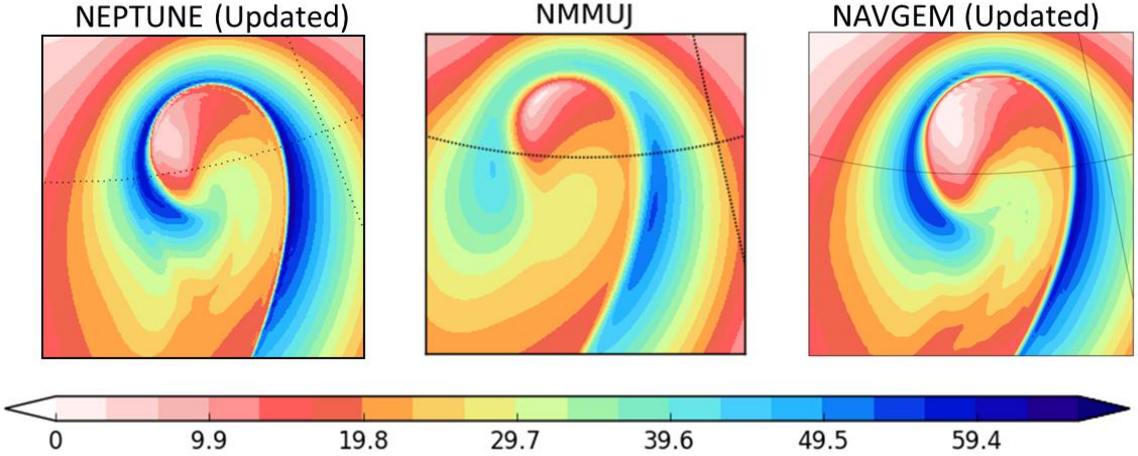
2-D decomposition with 2-TL dynamics
is not just more scalable, but
faster than 1-D

Dynamic Core Validation

Idealized Baroclinic Instability Test



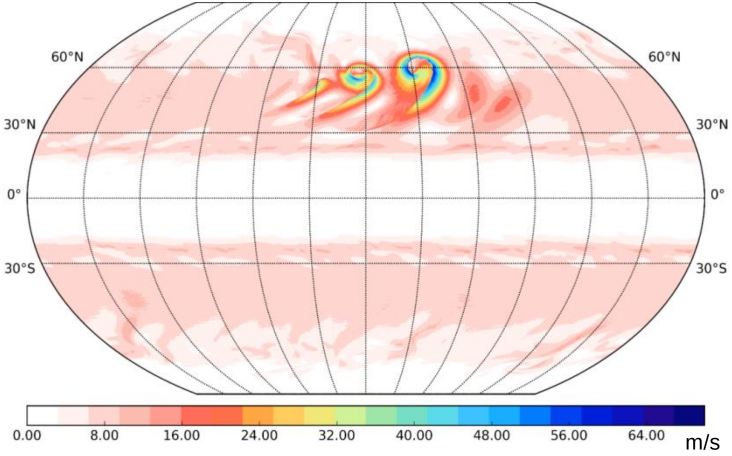
New NAVGEM idealized tests compare well to other dynamic cores evaluated by NGGPS

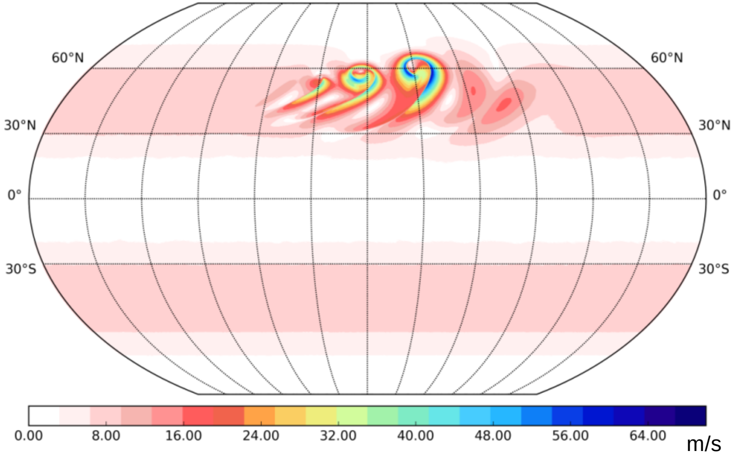


Dynamic Core Validation

Idealized Baroclinic Instability Test

- New recursive algorithm used to calculate Legendre Polynomials.
- New polynomials are up to 200 orders of magnitude different than legacy polynomials for high wavenumbers.
  - Legacy = O(-13) above wave number 300
  - New = O(-210) above wave number 300



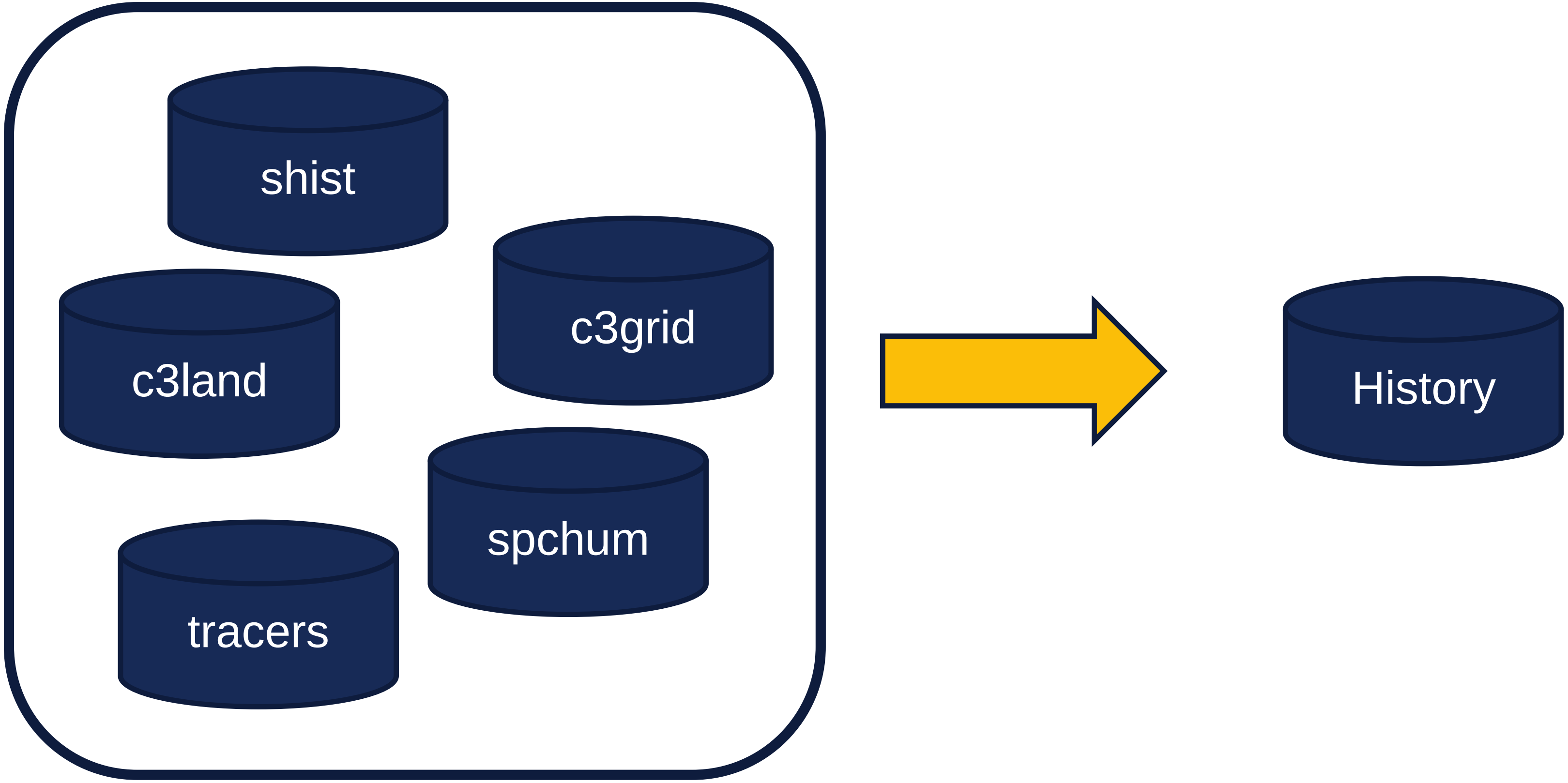


Result: Significantly less noise in solution!

Asynchronous I/O

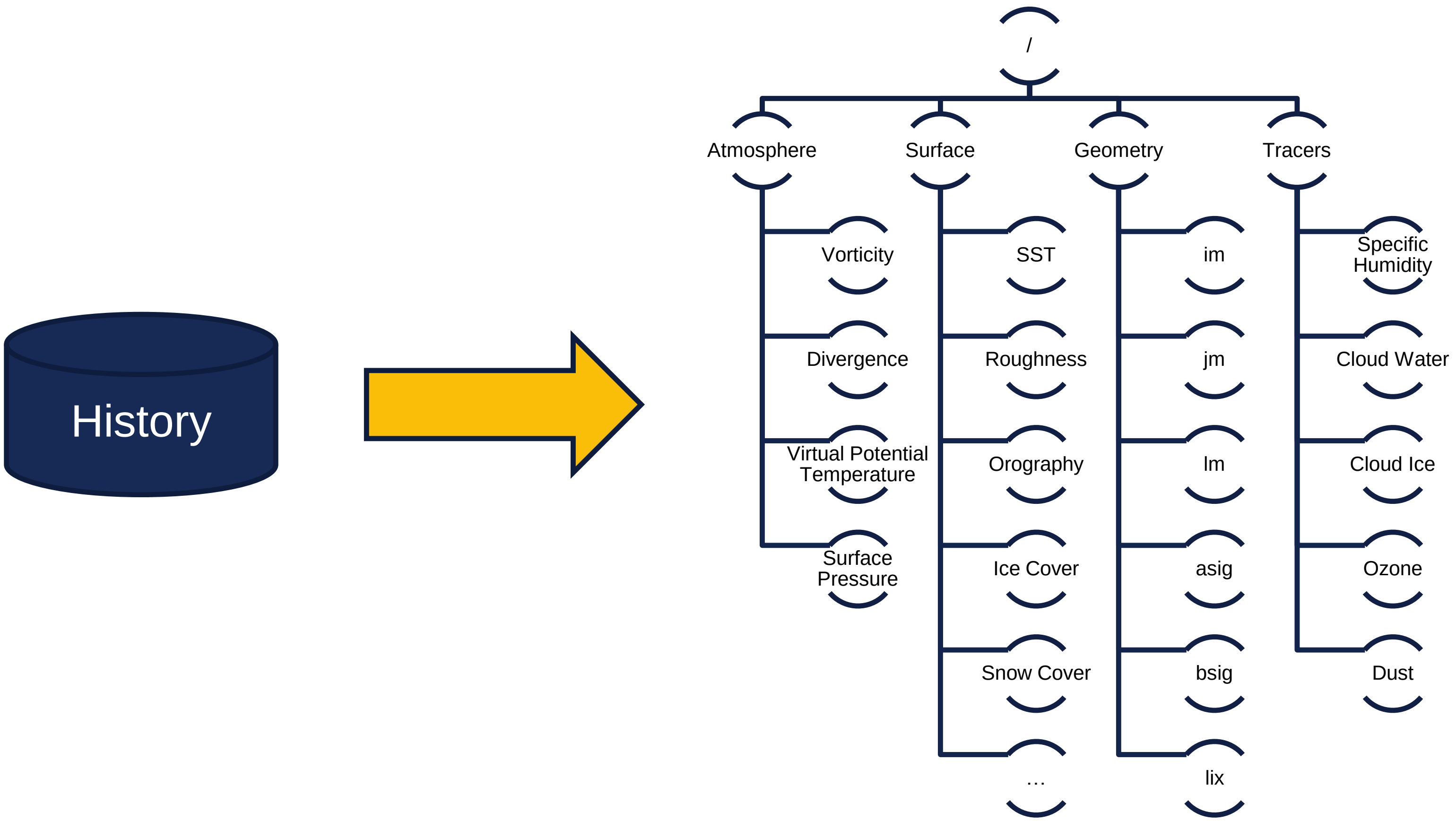
U.S. NAVAL RESEARCH LABORATORY

New I/O Method uses Parallel HDF5 file format



Fortran Binary, Limited Metadata

HDF5, Rich Metadata



Much easier file format to work with than before, but we are seeing performance impacts at high core counts (disproportionately expensive I/O).

Default Configuration – 16 nodes for forecast model + I/O

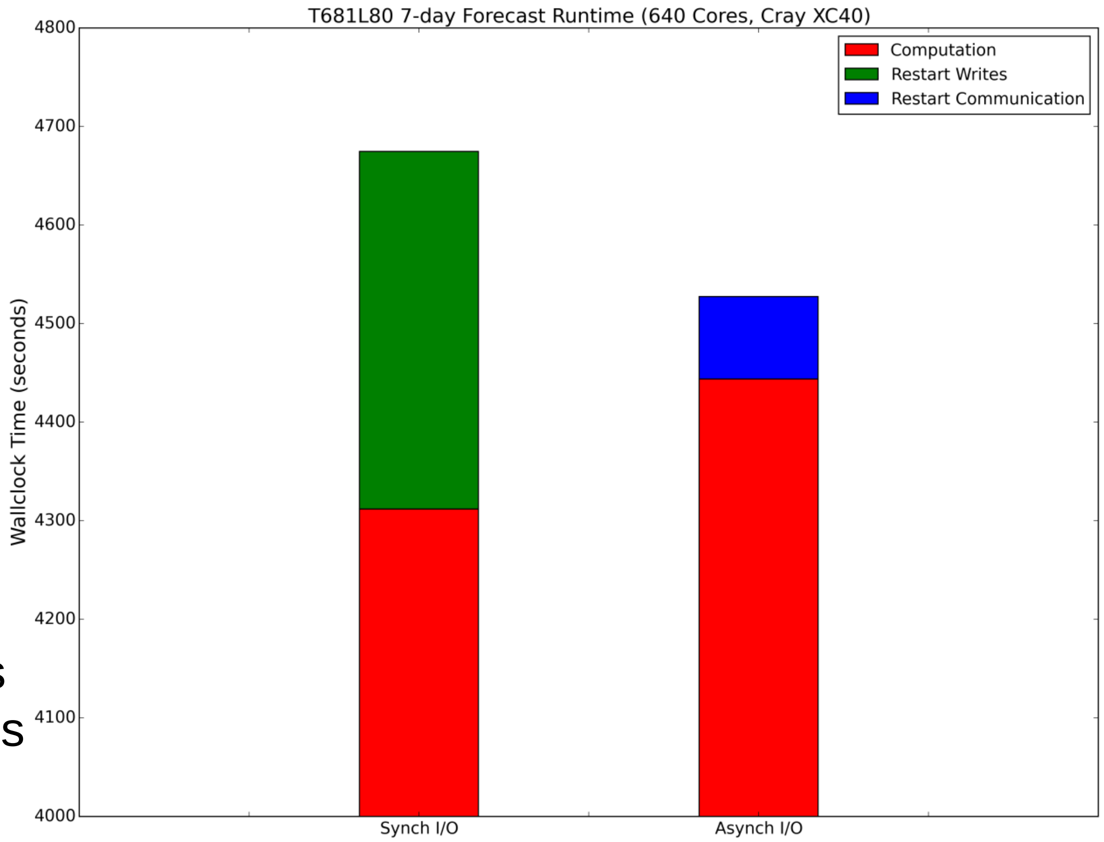
Asynchronous Configuration – 15 nodes for forecast model + 1 node for I/O

Future Asynchronous Configuration – 16 nodes for forecast model + 1 node for I/O

Use of Earth System Modeling Framework (ESMF) library takes care of mapping the distributed grids between the model component cores and the I/O component cores (as opposed to explicitly creating a separate MPI process pool to handle I/O within the forecast model)

In a fair timing comparison that kept the number of total cores the same, the additional cost of computation was offset by a much reduced I/O time.

To meet requirement to run on multiple DoD HPC platforms, this will allow us to tune I/O resources separately from compute resources

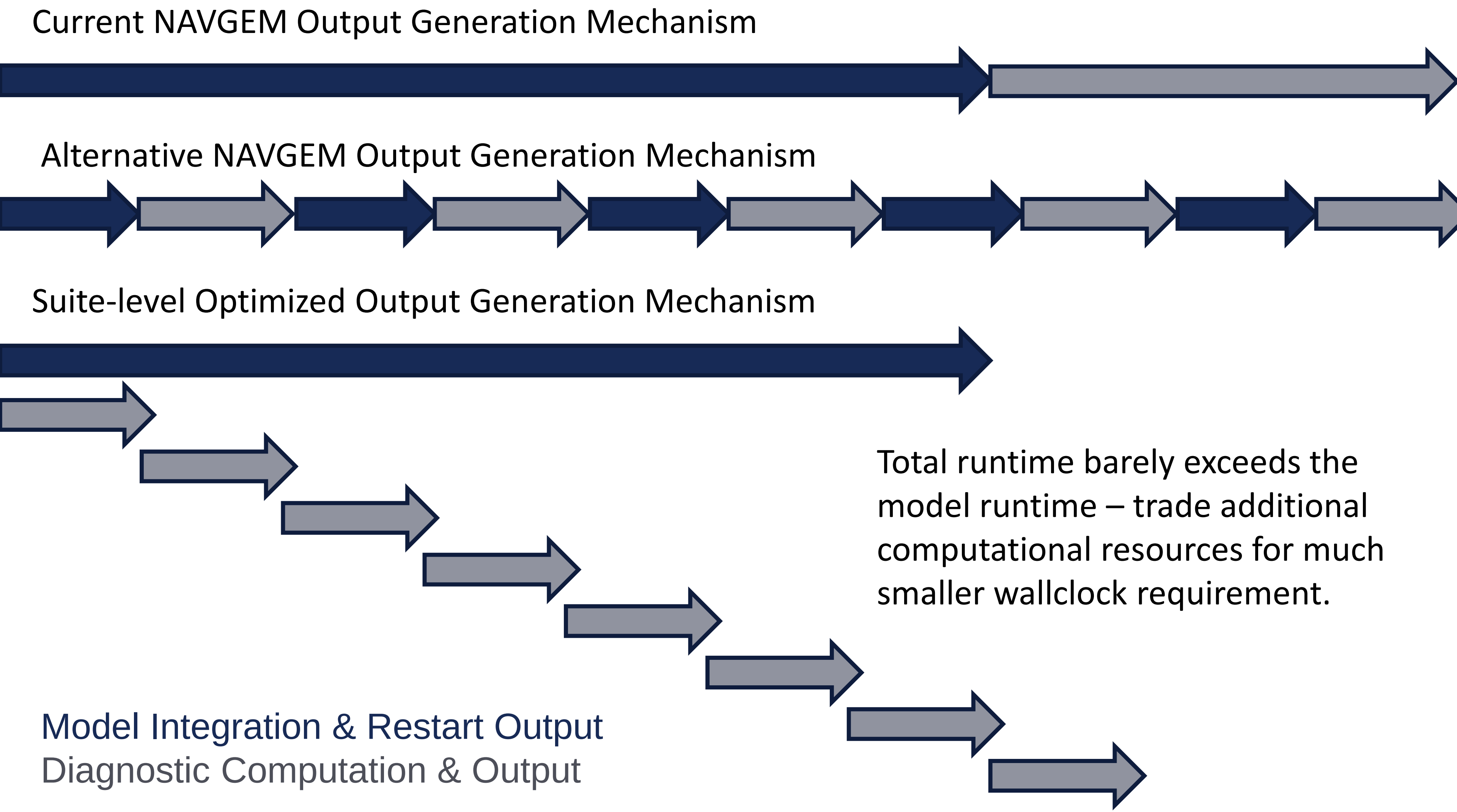


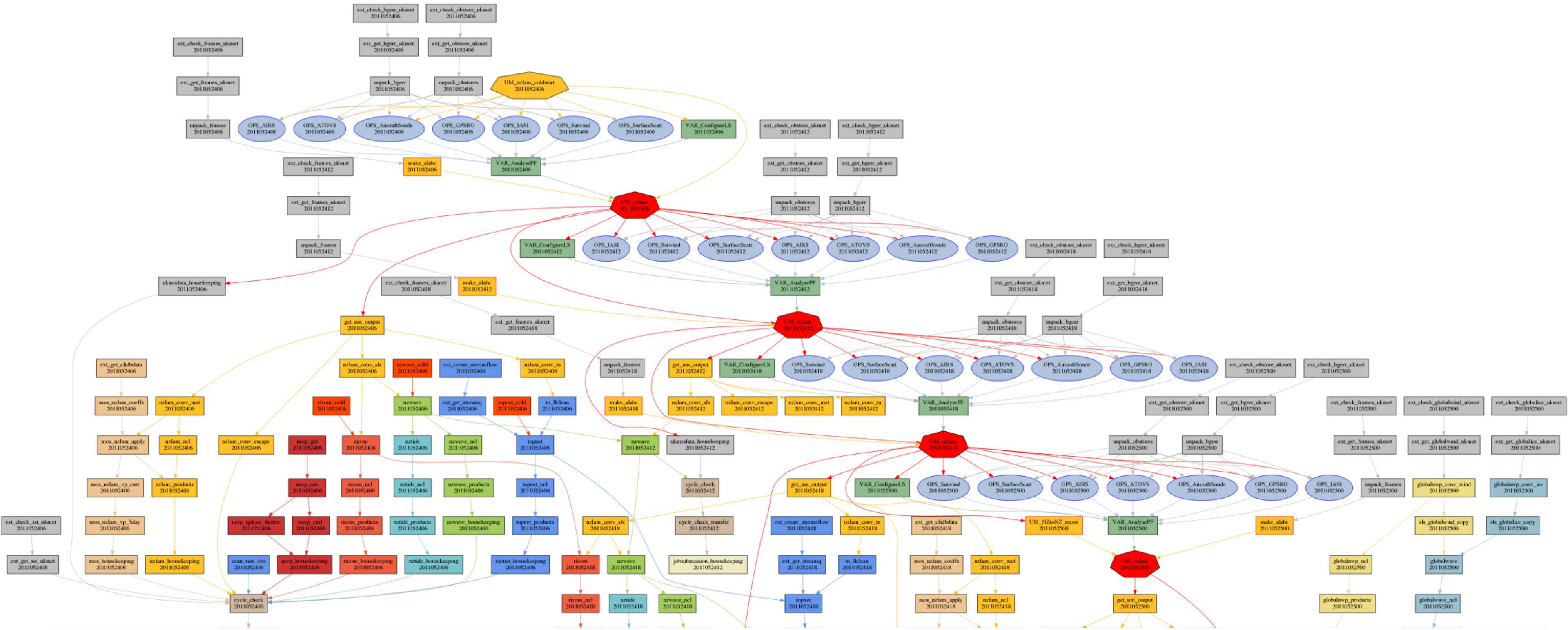
Multi-threaded ESMF should permit the computational cost to be similar between the two!

Asynchronous I/O

- Implementation via ESMF could potentially simplify the support multiple output streams – get the UKMO-style “I/O Server” without needing to manually manage MPI communicators
- Applicable to next-generation models as well – may be important to limit I/O when model is running over O(100,000) cores
- ESMF components in multiple threads hasn’t been tried before (because it’s a *really* bad idea for most ESMF applications)

Task Parallelism


Current NAVGEM Output Generation Mechanism
Alternative NAVGEM Output Generation Mechanism
Suite-level Optimized Output Generation Mechanism
Total runtime barely exceeds the model runtime – trade additional computational resources for much smaller wallclock requirement.
Model Integration & Restart Output
Diagnostic Computation & Output



Running a modeling system is more than just the forecast model – use workflow manager to efficiently overlap observation processing, data assimilation, forecast model integration, model post-processing, forecast sensitivity/observation impact, and transition seamlessly to (near-)real-time

Implications for Distributed Systems

- Two important tasks – job submission and job monitoring
- Job submission
  - Cylc supports remote job submission via SSH access to the remote host
  - For future (e.g. cloud-based) job submission, a web-services architecture may be an alternative
- Job Monitoring
  - When direct network access is possible, cylc can receive notifications from running tasks
  - Cylc supports SSH-based polling to remote job hosts if connection back to the main server process is unavailable
  - Recent cylc updates introduced an HTTP(S) based system to communicate back to cylc server: may be able to also modify polling infrastructure to use web services

Implications for Distributed Systems

- Workflow managers include their own implementation of a message queuing system
- Cloud services (e.g. Amazon) include implementations of message queues/notification services
  - Amazon SQS (Simple Queue Service) allows processes that can’t see each other directly to access cloud-based message queue
  - For typical NWP workloads, message service costs are quite reasonable (<1M requests/month are free)
- Cloud platforms may not be the best choice for operational NWP, but may serve as a useful middleman for managing distributed resources elsewhere!

Upcoming Efforts

U.S. NAVAL RESEARCH LABORATORY

Upcoming Efforts

- Now that the MPI decomposition is squared away, focus on OpenMP-based parallelization
- Incorporation of optimized radiation component – perhaps as separate ESMF component?
- Investigate multithreaded ESMF approach for atmosphere component with I/O – advantageous in the coupled system and may be applicable to ocean (HYCOM) and sea ice (CICE) as well
- Incorporation of more metadata in model files – can capture information about model versions, configuration, etc. (especially important for extended-range runs)
- Performance characterization of fully-coupled system

Significant Upgrades to Navy NWP

- New 2-D MPI Domain Decomposition
  - Vastly improves scalability over what adequately served for many years
- Asynchronous I/O
  - ESMF Infrastructure may permit easy adoption of multiple output streams for scalable I/O in earth system models
- Task Parallelism
  - cylc workflow manager has been an incredibly beneficial change to research modeling infrastructure