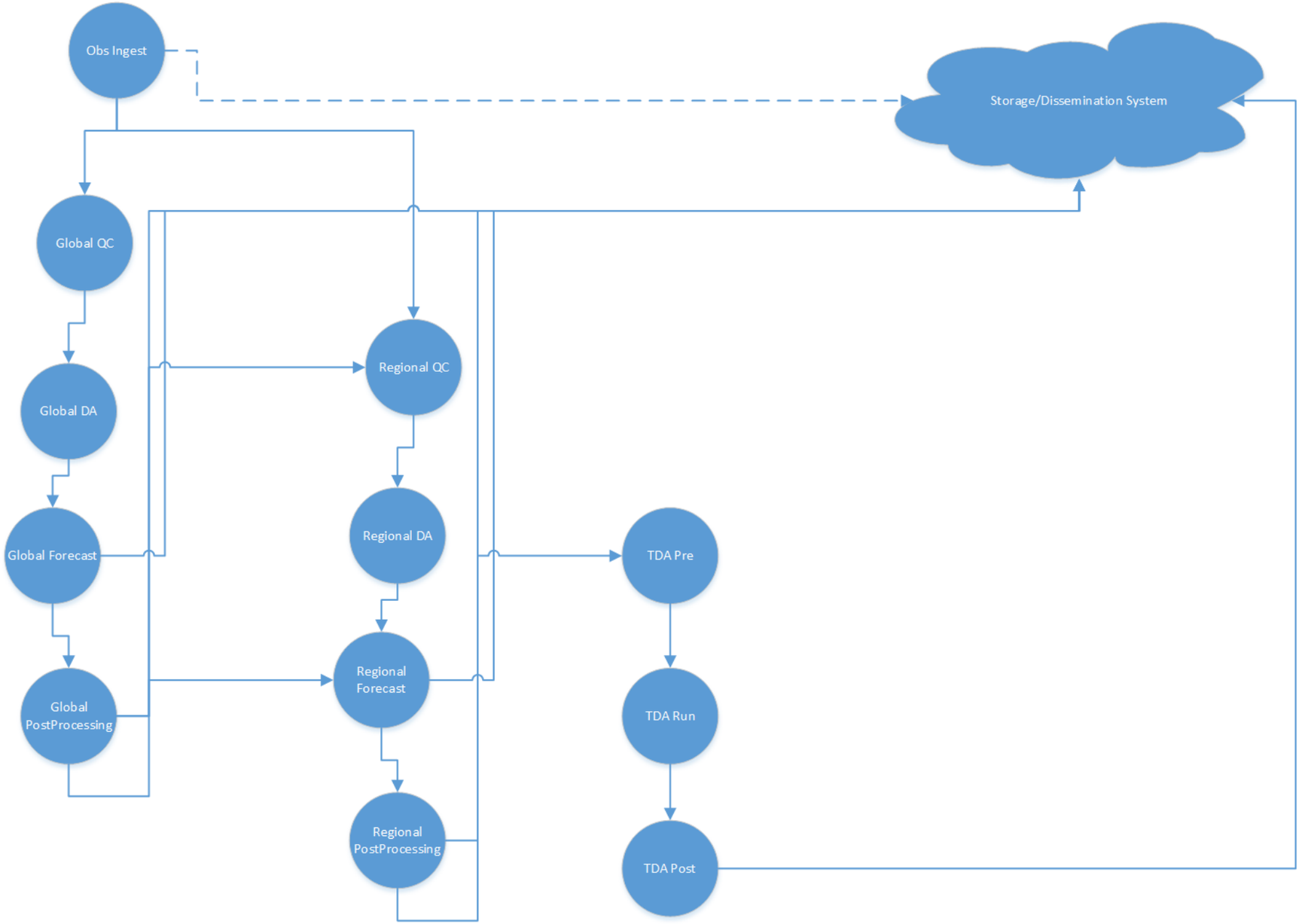
U.S. NAVAL RESEARCH LABORATORY

I/O Challenges for U.S. Navy METOC Modeling: From Data to Decisions

Timothy Whitcomb
NRL Marine Meteorology Division
September 2019

Typical interconnected workflows – from trusted observations to trusted products downstream



This workflow is repeated for multiple systems (e.g. atmosphere and ocean).

ESPC IOC

1 Day of ESPC Cycling Using cylc Workflow Manager

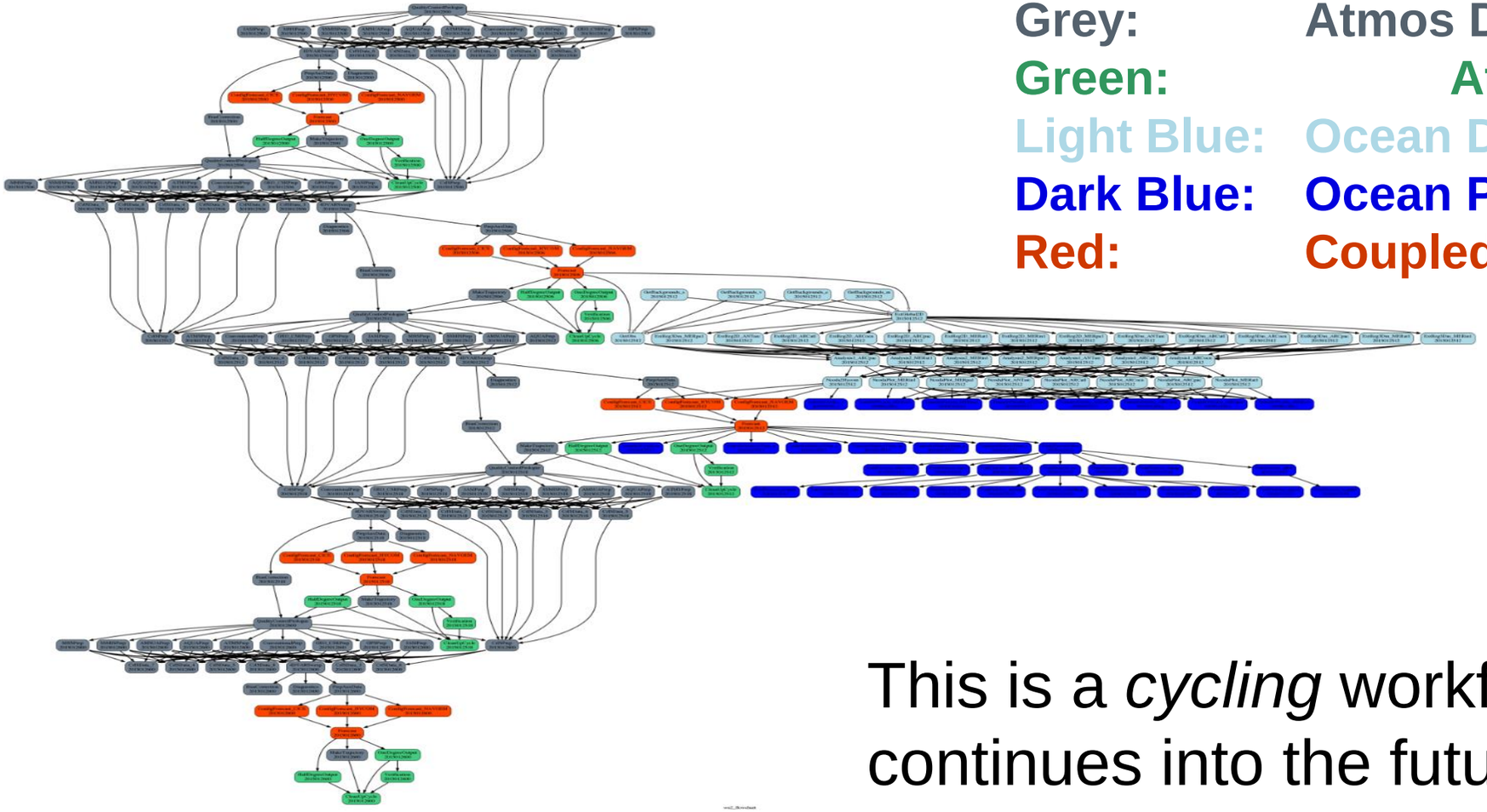
00Z

06Z

12Z

18Z

00Z



Grey: Atmos DA

Green: Atmos Post Process

Light Blue: Ocean DA

Dark Blue: Ocean Post Process

Red: Coupled Model Tasks

This is a *cycling* workflow that continues into the future.

Application-Centric View

Single Executable

Multiple Executables

Monolithic Tasks

Single-purpose small Tasks

Workflow-Centric View

Application-Centric View

Single Executable

Multiple Executables

Monolithic Tasks

Single-purpose small Tasks

Workflow-Centric View

Inter-Application Data Exchange

Monolithic Tasks → Parallel filesystem, Fast on-node storage

Single-app Tasks → Parallel filesystem, ~~Fast on-node storage~~

I/O in workflows is analogous to MPI in applications (Bill Williams, 2019).
Is fast, on-node I/O between applications in a single workflow job analogous to OpenMP in applications?

System-Specific Comments

U.S. NAVAL RESEARCH LABORATORY

Observation Ingest

- Want indexed, high-speed access that supports eventual consistency, assured access
- Current bottleneck – many, many files (esp. for smaller satellite granule sizes)
- Need something platform-independent, able to be distributed

Global & Regional QC

- For global system, rely on task-parallelism to process conventional & satellite data quickly
- I/O for satellite data is a challenge as files get larger
- I/O for conventional data is a problem because of the wide variety of files involved
- For regional systems, need to support forward-deployed applications, supporting not only standard QC but also in-situ observations.

Data Assimilation & Forecast

- Computational Characteristics of data assimilation solver are different than the forecast model
- Different shapes (i.e. rectangles in a utilization chart) should be in different tasks?
- One solution – combine DA and Forecast into a single executable and pass data in memory. However, this makes it more difficult to balance different shapes
- Alternatively – NVRAM to give more operational flexibility via supporting fine-grained tasks?

Post-Processing

- Includes interpolation of model-space variables to people-space variables (e.g. evenly spaced lat/lon grids, pressure/height level data) as well as "post-processing" (e.g. bias correction, MOS, etc.)
- Running as part of the forecast model reduces I/O burden (for communicating state) but makes the overall system *less flexible* and further burdens the time-to-solution

Image Generation

- Changes often, need a way to quickly access new data
- Wanted: streaming data paradigm supporting image generation and processing as upstream data becomes available

Decision Aids

- Transform model state into more useful, but more esoteric forms
- Ultimately changes in forecast skill must be mapped to changes in decisions

Inquiry

U.S. NAVAL RESEARCH LABORATORY

Asynchronous I/O

- I/O Server (asynchronous I/O) has been very helpful for minimizing impact on forecast runtime by I/O. **However, we have only ever used this internal to an application.**

With NVRAM storage, can we have a *persistent* I/O server that allows high-speed I/O for passing data between tasks without the restriction of consolidating tasks and processes?

Global/Regional Links

- Need flexibility in terms of boundary conditions – new domains may be needed, so decouple boundary condition generation from the global model

Can high-speed, semi-persistent storage be leveraged to optimize boundary condition generation for regional models?

Tactical Decision Aids

- Often TDAs and uncertainty involve a combinatorial explosion: each individual execution may be fast, but quickly grows in requiring more and more (e.g. for radar propagation, number of azimuths * number of radii * number of elevations * number of ensemble members)

Again, can nonvolatile RAM storage provide a high-speed data source to knock down TDA runtime by supplying a high-availability, high-speed state?

The Main Issue

Can next-generation I/O with NVRAM storage allow maximum decomposition of tasks for greater flexibility, but maintain the performance of larger, monolithic tasks within a workflow?