Optimizing Data Flow and Availability

M.C. Macduff and K.J. Creel
Pacific Northwest National Laboratory
Richland, Washington

R.C. Eagan
Argonne National Laboratory
Argonne, Illinois

Introduction

As of March 2003, the data from all instrumented Atmospheric Research Measurement (ARM) sites is delivered to the Data Management Facility (DMF), processed for archival and made available for data quality and other internal ARM uses, all on an hourly basis. Processed and raw data are transfered to the Archive on a daily basis. This has resulted in a significant improvement in availability of processed and quality-checked data. The change in data delivery and processing methods was driven by problems managing distributed processing, multiple data paths to the Archive, delays in data flow and delays in quality checks of data. A combination of technical solutions converged to resolve these problems. By Fall 2002 the site data systems at all sites were running the same software and there was internet connectivity to the Tropical Western Pacific (TWP) sites. It was now possible to change the data flow. While limited bandwidth still placed limits on data transfer efficency, the transition was completed by updating the software controlling data flow on the data systems. The result is that within 20 minutes of being collected, the data from all sites is on the DMF. Within the next hour the data is processed into netcdf files. At this point, the data is available for quality checking or other internal use. At the end of the day raw and processed data are ready for transfer to the Archive. Streamlining the data flow has positively impacted operations and engineering support which contributes to improved data quality, data availability and ultimately, usefulness to the Science community.

Background

The initial deployments of each of the Southern Great Plains (SGP), TWP, and North Slope of Alaska (NSA) Atmospheric Climate Research Facility (ACRF) sites, for various reasons, used slightly different hardware and software designs. The sites also differed (and likely always will) in their accessibility and on-site staffing. This situation required a set of developers to support each site. The fact that the TWP sites did not have network bandwidth sufficient for transmitting data required unique procedures for data transfer and processing. Because of some of the difficulties in maintaining the sites at TWP and NSA, the DMF at Pacific Northwest National Laboratory (PNNL) was created to serve as a collation point for the data. Any changes required to update an ingest could be easily fixed and installed on the DMF. The affected data from the site could then be reprocessed. Until the software could be deployed to the site, the DMF would continue to reprocess that data stream prior to shipment to the Archive. This was all very manual and significantly impacted timeliness and reliability of data flow.

Data System

The data system had been based on Sun Microsystems Solaris Operating Environment at all of the sites. Based on experiences of the SGP site using Linux at the extended facilities (EFs), a Linux collection machine was added to the data systems at all measurement locations. Through a couple iterations of hardware, all of the sites are using the same equipment. This was a significant benefit but the real effort involved the software. Managing 3 different ingests for the same instrument at 3 sites was a recipe for confusion. To effectively manage this a single software development environment that allowed the developers to create a single ingest for an instrument for all sites was developed. The final major effort was to upgrade all of the sites to run identical hardware and software.

Network

While the SGP and NSA sites have had reliable network connectivity since their inception, the TWP sites location presented more of a challenge for shipping data to the DMF.

Recently, data collected at the Manus and Nauru sites were transferred to the DMF via removable hard drive media, periodically removed and air freighted by local support staff. This transfer process often took several weeks.

In order to have a daily assessment of the proper functionality of the measurement and collection systems at Manus and Nauru, a method of shipping hourly status messages via geostationary operational environmental satellite (GOES) satellite to the DMF was developed and implemented. The messages were limited to a few hundred bytes, so it was necessary to be clever about selecting what information was sent.

In early 2002 the ARM program identified a satellite service provider that could support ground stations at the ARM Pacific sites and was reasonably priced. Since the majority of the ARM program traffic was shipping data off-site, systems that provided 256Kbps uplink and 64Kbps downlink were installed at Manus and Nauru. These sites now had continuous network connectivity with sufficient bandwidth to transfer virtually all data over the network link in near real-time. The Darwin site has been connected to the internet since installation.

At this point all measurement sites had continuous network connectivity.

Data Flow

The overall process of getting the data from the instrument to the Archive creates a data flow. As each site changed, the data flow necessarily changed. The original design of our systems was based on hourly collections of data on site and daily deliveries between sites. The SGP shipped data daily to the Archive and performed only daily shipments of process data to the DMF as necessary to support Value

Added Products. The NSA and TWP, however shipped data daily to the DMF. To avoid sending incomplete daily data sets to the Archive, transfers to the Archive were two or three days later than the day of collection. Thus it could be several days before users or even ARM infrastructure could have access to processed data.

The Data Quality Office (DQO), a key user in the infrastructure, and a key component in the data quality control process, developed tools and procedures to analyze the quality of the data streams. The initial implementation of the DQO was at the SGP site. The DQO depended on timely (preferably hourly) updates to the datastreams. The delayed delivery of TWP and NSA (transferred from DMF to SGP) data limited the operational benefit to those sites. All of these factors led to a review of the data flow within ARM. In the Fall of 2002, it was recognized that re-architecting the data flow could significantly improve the ability to automate the process and subsequently improve the timeliness of data availability and efficiency of operation.

The New Design

A result of the design review was to implement the data flow such that the official ingests for all datastreams, including SGP, run at the DMF. This reduced the load on the network capacity to all of the sites by reducing the off-site data flow to only shipping hourly raw data and only to the DMF (see Figure 1). The second key component of the new data flow architecture is the relocation of the DQO to the DMF.

Under this architecture the DMF could then process all raw data hourly and make the products immediately available to the DQO and other internal users. Data could be shipped to the Archive the next day, thus providing equal timeliness of all data products from all sites.

The new design resolved the known issues by co-locating the DQO service with the DMF and only transmitting the raw data across the network. Converting the DMF to hourly processing eliminated problems stemming from the previous daily flow. The solution provided added benefit as a central location for real-time access for mentors and IOP participants. The new stream-lined data flow also simplified the architecture and thus the management of our data systems (see Figure 2).

Implementation

In order to actually change the data flow it was necessary to update some of our core software to accommodate the changes. These core pieces of our data flow have been proven over time and it was necessary to exercise great care in designing and coding changes to them. Elaborate tests were setup to validate the changes. Because of the internals of our processing system, it is necessary to process the data in-order. On a local system, this is nearly a trivial problem. Across long-distance networks, it very possible to have files delivered out of order. Therefore, specific code and procedures were developed to ensure in-order delivery of data from the collection on site to processing at the DMF.

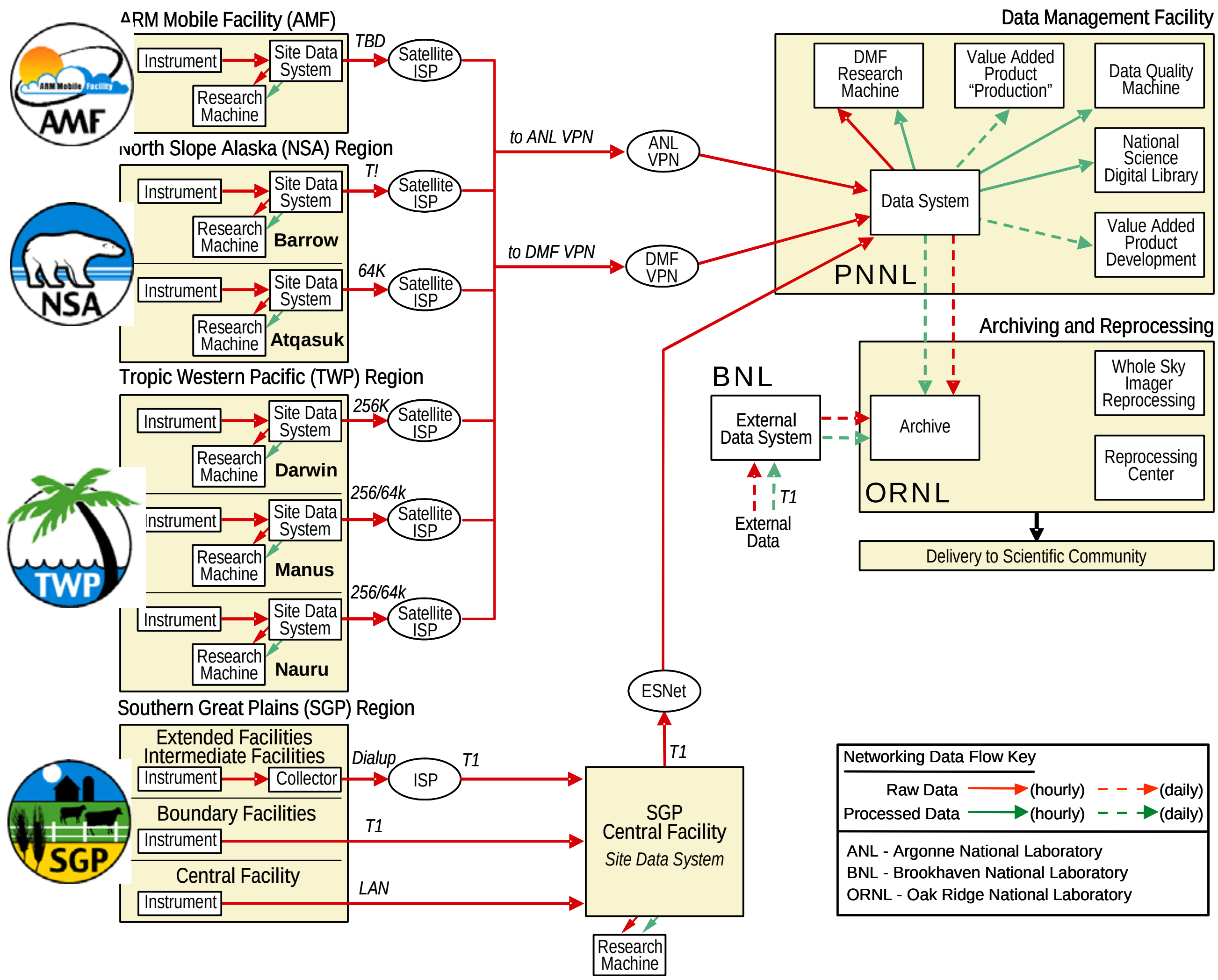

Figure 1. The new hourly data flow from the sites.

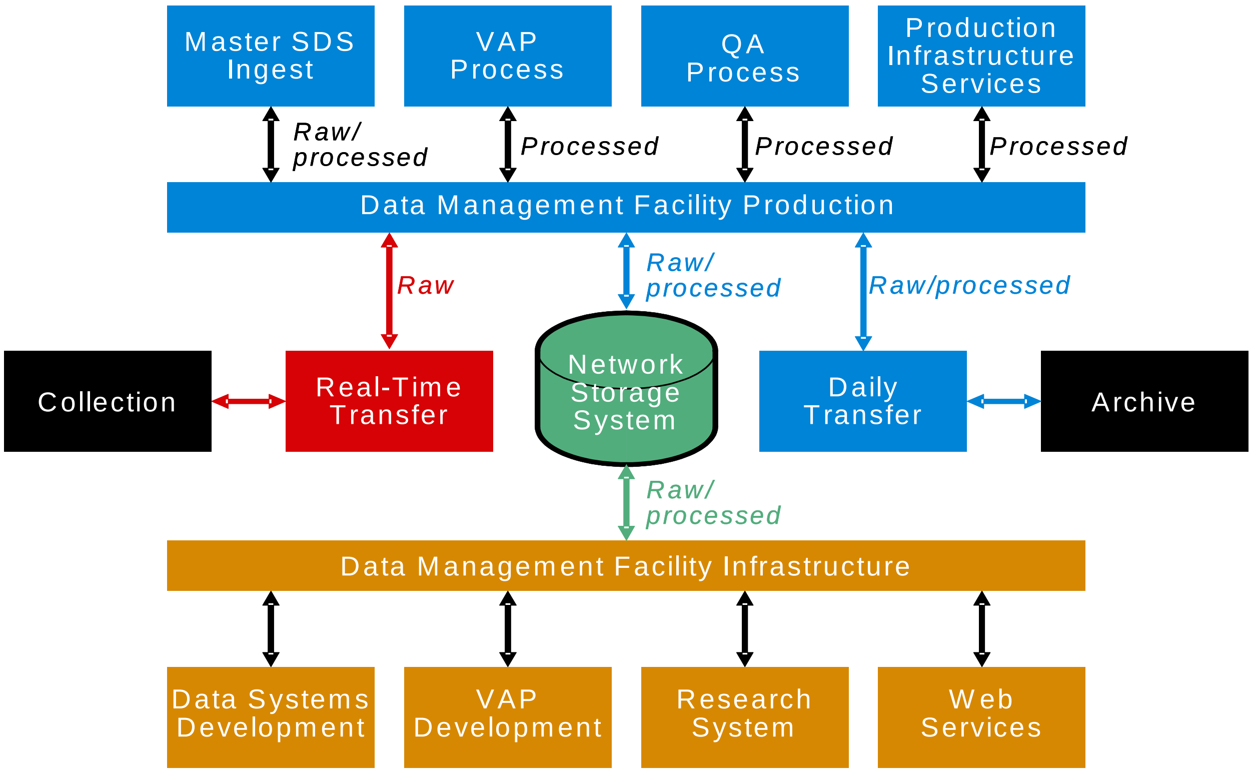

Figure 2. The DMF's internal data flow.

Doorstep

In the current architecture, the on-site collection system hands the hourly data to the on-site processing system via a "doorstep" process. The doorstep process has been modified from the original implementation to replicate the data. Thus, one copy continues to the local system for processing and local viewing, and the other copy is shipped off-site. This was implemented with a fairly simple algorithm that will guarantee that either both copies exist or none.

Site_Transfer

Site_transfer has been the work-horse for inter-site data delivery on ARM since ARM's beginning. It does the work of moving the copy of the data from the doorstep to the DMF. Delivering data in-order was a straight-forward change of sorting with timestamps. However, this introduces the problem of a large instrument inserting a glut of data. This would effectively block all data from that site until the backlog was transmitted. Therefore, the idea of priorities for different data types (IE logs always go first) was developed. To accommodate the complex structure of many small files from the doorstep it was necessary to transfer a directory tree as well as files between sites.

Solving these problems within doorstep and site_transfer made the new data flow design possible.

Performance

After the initial installation of the new data flow design to the TWP and NSA sites, it was observed that data flow did not achieve the throughput expected. In fact at Barrow, the system was not able to keep up with the WSI data, this issue had not previously been a problem. Understanding the performance problem involved a subtlety of networking: satellites have long latency. This meant that a 256k line with 1 sec round trip, could hold 64KB “in the air”. The average file size for hourly transfer was ~1KB. Since we use ftp, and sent the files in-order serially, we were only using a small portion of our bandwidth.

The solution to the performance problem involved the realization that in-order delivery is only required for each instrument. Using the beginning tags on the files, the files could be sorted by instrument and by time. Site_transfer was further modified, using threads, to open multiple ftp sessions and deliver files in parallel. Using this technique, throughput improved to the extent the process can now easily handle backlogs of any kind from any site.

Deployment

The deployment was completed over a 1 month period as each site was incrementally converted to the new paradigm. Due to the work done with upgrading the data systems and networks, the process of radically changing the data flow was a simple procedure of updating a few software packages. The system was monitored very carefully during the transition and virtually no problems were observed. The design worked exactly as expected.

Conclusion

As a result of several multi-year tasks to improve the data systems and networking of the ACRF sites, it was possible to implement a significant change in the data flow. The current data flow architecture has not only made processed, quality checked datastreams available to scientists and other users in near real-time but also has enabled the ACRF infrastructure to manage and monitor the process in a much more efficient manner.