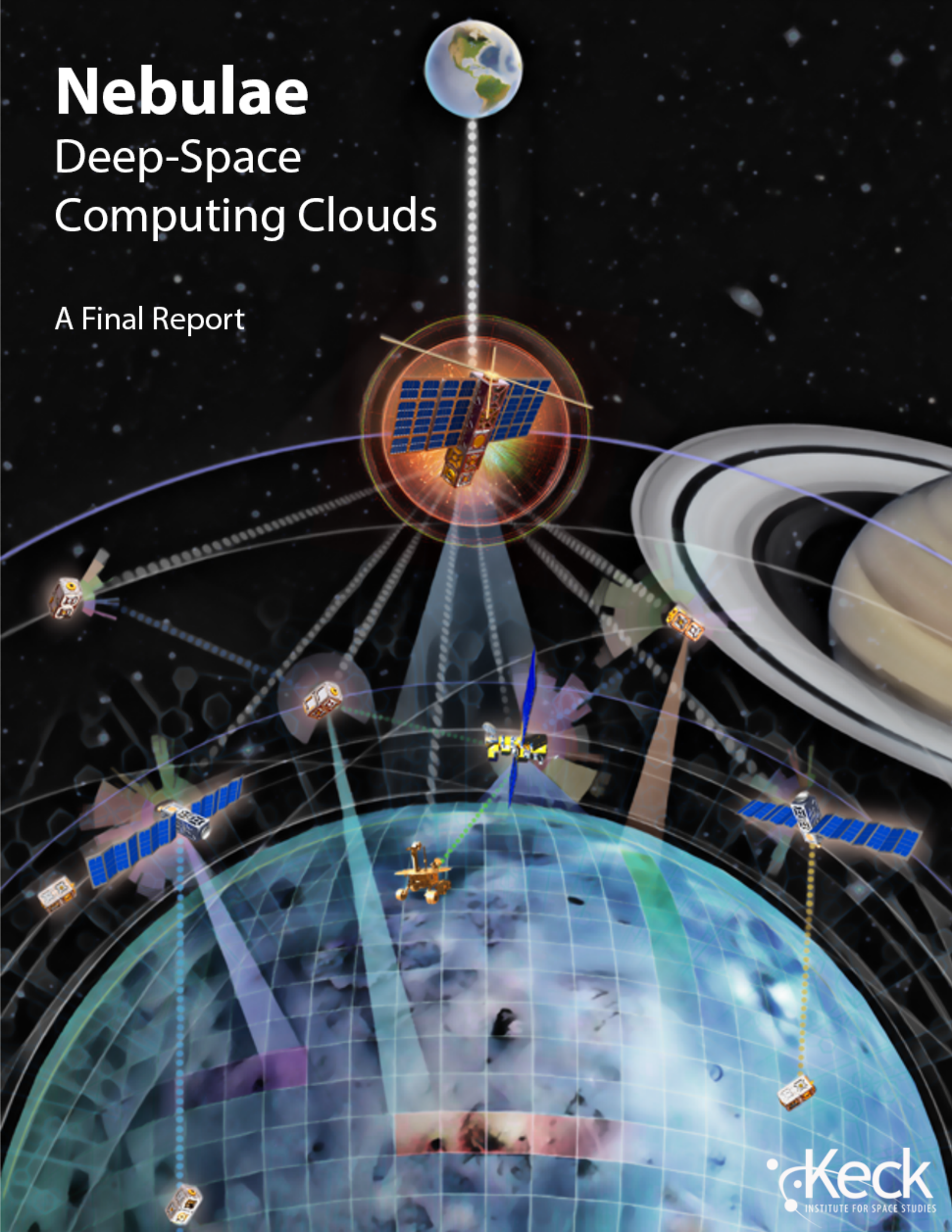
# Nebulae

## Deep-Space Computing Clouds

A Final Report



Keck Institute for Space Studies

**Nebulae:**
**Deep-Space Computing Clouds**

Study Report prepared for the W. M. Keck Institute for Space Studies (KISS)
Study Start Date: August 26, 2019
Study End Date: September 4, 2020

Team Leads:
Joshua Vander Hook / JPL
Rich Doyle / JPL
Valerie Fox / University of Minnesota
David English / Lockheed Martin
Ashish Mahabal / Caltech

Director: Prof. Tom Prince
Executive Director: Michele Judd
Editing and Formatting: Meg Rosenburg
Cover Image: Chuck Carter/Keck Institute for Space Studies (KISS)
Header images: NASA, NASA/JHUAPL/SwRI, NASA/JPL-Caltech, NASA/JHUAPL/Carnegie Institution of Washington, NASA/JPL-Caltech, NASA/JPL/Space Science Institute, NASA/JPL-Caltech/SwRI/MSSS/Björn Jónsson, NASA/JPL/Space Science Institute, NASA/JHUAPL/SwRI, NASA/JPL/Space Science Institute, NASA/JPL-Caltech/SETI Institute, NASA, NASA

# Study Participants

**August 26–30, 2019**

Leon Alkalai
Jet Propulsion Laboratory

Dmitriy Bekker
Johns Hopkins Applied Physics Lab

Morgan Cable
Jet Propulsion Laboratory

Julie Castillo
Jet Propulsion Laboratory

Alice Cocoros
Johns Hopkins Applied Physics Lab

Les Deutsch
Jet Propulsion Laboratory

James Dickson
Caltech

Andrew Dittrich
USAF

Richard Doyle
Jet Propulsion Laboratory

David English
Lockheed Martin Space

Valerie Fox
Caltech

Eric Frew
University of Colorado

Joseph Goldfrank
Stanford University

Trent Hare
USGS

Shayn Hawthorne
Amazon Web Services

Jason Hofgartner
Jet Propulsion Laboratory

Robert Hood
JASRC Federal/NASA Ames

Ashish Mahabal
Caltech

Lukas Mandrake
Jet Propulsion Laboratory

Sreeja Nag
NASA Ames / BAERI

Mario Parente
University of Massachusetts

Raphael Some
Jet Propulsion Laboratory

David R. Thompson
Jet Propulsion Laboratory

Jason Tichy
NVIDIA

Joshua Vander Hook
Jet Propulsion Laboratory

**August 31–September 4, 2020**

Leon Alkalai
Jet Propulsion Laboratory

Morgan Cable
Jet Propulsion Laboratory

Ken Center
Orbit Logic

Les Deutsch
Jet Propulsion Laboratory

Richard Doyle
Jet Propulsion Laboratory

David English
Lockheed Martin Space

Valerie Fox
Caltech

Anthony Freeman
Jet Propulsion Laboratory

Joseph Goldfrank
Stanford University

James Gosling
Amazon Web Services

Gregg Hallinan
Caltech

Trent Hare
USGS

Bob Hodson
NASA Engineering Safety Center

Robert Hood
JASRC Federal/NASA Ames

Mike Johnson
NASA Goddard Space Flight Center

Joseph Lazio
Jet Propulsion Laboratory

Ashish Mahabal
Caltech

Lukas Mandrake
Jet Propulsion Laboratory

Ben March
University of North Dakota

Sandeep Prasad Chinchali
Stanford University

Ray Ramadorai
Blue Origin

Rob Ruyak
Amazon Web Services

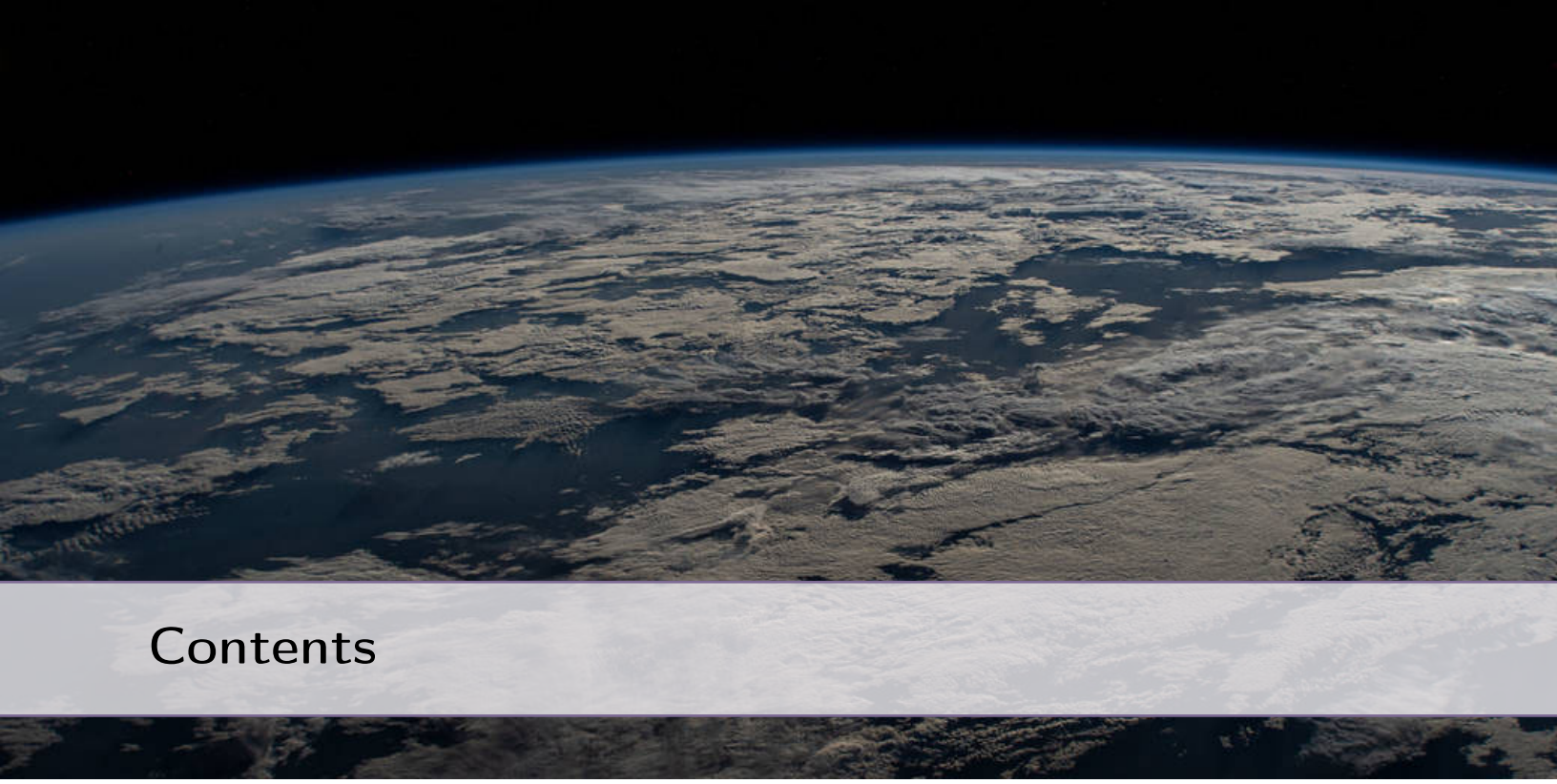
Daryl Schuck
World Wide Public Sector

David R. Thompson
Jet Propulsion Laboratory

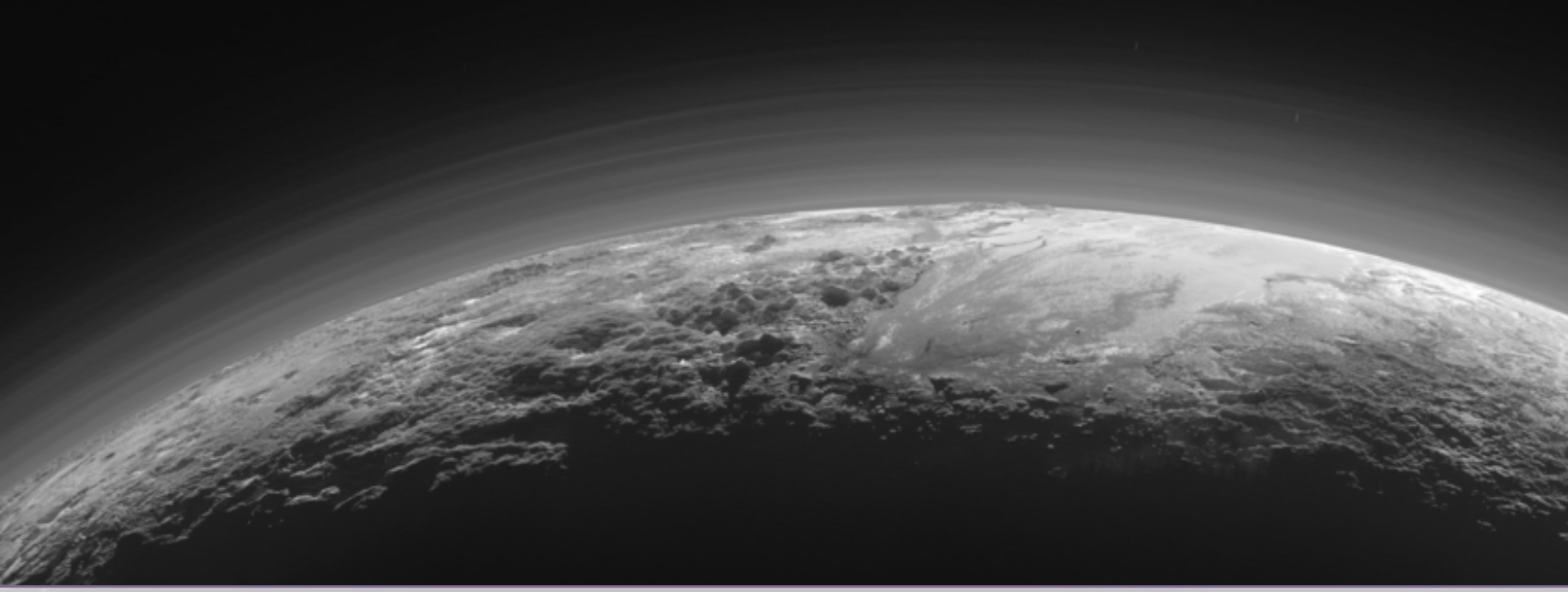
Joshua Vander Hook
Jet Propulsion Laboratory

Kiley L. Yeakel
Johns Hopkins Applied Physics Lab

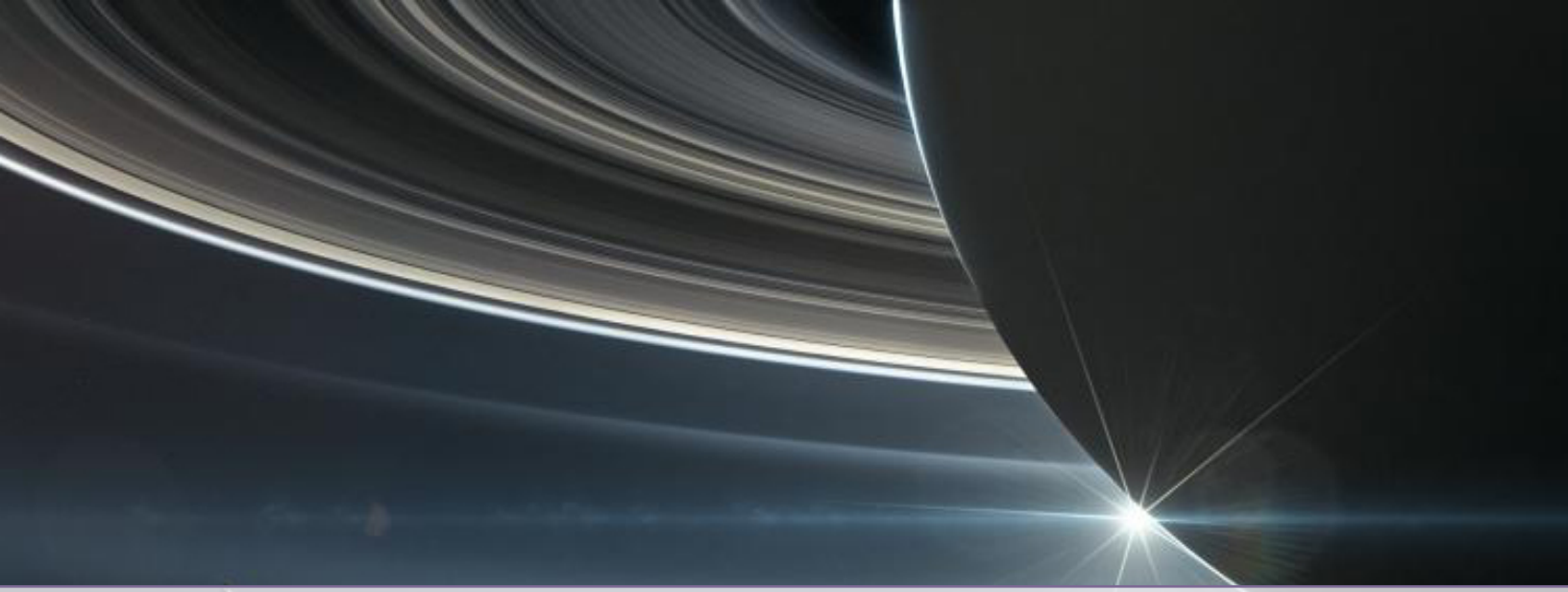
Kate Zimmerman
Amazon Web Services



# Contents



# List of Figures



# Executive Summary

Science exploration is about extending practical reach and intellectual assessment to increasingly remote and operationally challenging environments. Historically, as a pragmatic consideration, each space mission has been approached as a self-contained endeavor, with specific science objectives and a well defined resource envelope within which to accomplish those objectives. This paradigm has served well, particularly for the unprecedented risk management challenges that attend first-of-a-kind exploration activities.

However, challenges of efficiency and sustainability inevitably appear, calling for new ideas on how space missions can build on each other, not only in terms of science yield and understanding, but through incrementally deployed services and infrastructure, such that each new endeavor need not self-carry all the required resources.

This thinking is the inspiration behind the Nebulae set of concepts for deploying computing, data storage, networking, and cloud services as infrastructure to remote regions of the solar system, in robust and scalable fashion.

Nebulae does not represent a single concept. We unpack the idea into multiple forms: 1) Vigilant Instruments, pushing on today's concepts for onboard science data analysis; 2) Data Server in the Sky, to be as reliable and trusted as any ground-based archive; 3) Observing System in the Sky, consonant with continuous spatial and temporal observing coverage of Earth as a System; and 4) the complementary Data Cycler for physically returning remotely acquired data to Earth in ongoing if not immediate fashion, for additional completeness and robustness.

We examine several use cases, ranging from Mars (where the advantages of having multiple platforms active concurrently, e.g., for relay operations, is already well demonstrated), to astrophysics (for the robust capture of relevant detections), to Earth as a special case (given that the communications challenges of deep space do not apply).

We also ask the retrospective What-If? question: "What additional value-added science might historical and/or current space missions have accomplished if Nebulae-style capability had been available to those missions?" Results of this exercise not only further illuminate the use cases, but could lead to risk-managed enhancements of flying missions, or concepts for new near-term technology demonstrations, even as we pursue the Nebulae concept itself as a more strategic opportunity.

Finally, we examine extant capabilities and emerging technologies for computing, data storage, networking, and cloud services, and project their availability for deployment within future Nebulae instantiations, describing the possibilities over the next few decades, to stimulate pragmatic excitement, and the beginning of strategic planning.

Our aim is to shine a light, from our perspective, that today's science mission exploration paradigm, having served us as a community in exemplary fashion, simply will not scale as the reach of our exploration extends further, as it must, into more remote, unknown, and fascinating environments. We offer new concepts to address that reality via a shift to a more sustainable exploration paradigm—Nebulae—within the spirit of community engagement and discussion.

# 1. Introduction

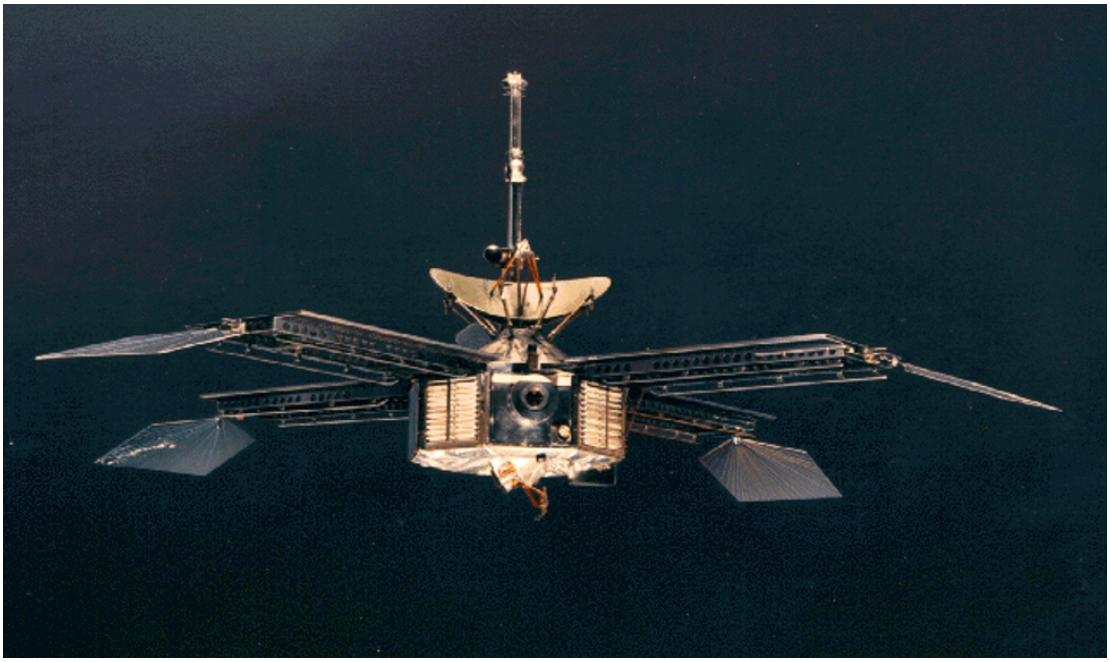


Figure 1.1: Mariner 4 spacecraft.

In 1964, the Mariner 4 spacecraft became the first mission to return an up-close image of another planet, Mars. At Mars, Mariner 4 captured imagery according to a predefined command sequence as it flew past the planet. Onboard processing converted each frame to a digital array of 40,000 pixels nearly instantaneously on board the spacecraft, and the resulting 260,000-bit digital data was sent to persistent storage drives onboard (around 20


Figure 1.2: Hand colorized image data returned by Mariner 4.

such images could be stored), and later transmitted to Earth at around 34 bits per second. Each image transmission must have taken over two hours. On Earth, as the digitized image was being processed, engineers in the Telecommunications section of the NASA Jet Propulsion Laboratory (impatiently, we can imagine) used art pastels to hand colorize the pixel values to verify the camera's functionality. A result is shown above in Figure 1.2. Perhaps even when colored by hand, the rendering was still faster than the communication from Mariner to Earth.

Despite a half century of technological progress, this operating paradigm has remained largely unchanged since it was introduced with Mariner 4, albeit with more pixels and fewer pastels. The primary role of a spacecraft is still to gather, temporarily cache, and transmit to Earth any data requested by Earth-side scientists.

This report does not advocate to replace that paradigm—it is trusted, expected, and highly productive. Instead, we attempt to augment it, using the impending availability of high throughput computing and large volume storage. We asked, *What can be done with those data we chose not to gather, not to store, or not to transmit to Earth?* This simple question branched out in many directions during the course of two workshops.

To summarize, we predict and advocate a new paradigm in which spacecraft with computing and storage nodes will generate more scientific value per mission by caching large volumes of data for purposes other than immediate downlink—such as providing data summaries, allowing remote access to and, eventually, performing iterative, human-directed analysis on

board. This more dynamic environment is enabled by increasing the onboard or networked processing and storage capacity of next-generation missions.

The computing-enhanced system can be accessed or tasked remotely to broaden the reach of scientific discovery, both as originally intended during mission design and serendipitously during operation or post-mission. The spacecraft begins to resemble a shared data resource. Since this is akin to the cloud-computing paradigm on Earth, yet in deep space, we dub the paradigm Nebulae after the deep-space clouds.[1] As has happened with Earth-side data centers, we anticipate that the introduction of data science techniques to deep space missions will enable real-time, autonomous decision making or summarization of the data gathered but not yet downlinked. The challenge is to improve the available onboard data storage and processing by orders of magnitude. While space capability lags on-Earth/commercial technology by 10–15 years typically (driven by radiation hardening and system qualification challenges), we are moving to a point where the new paradigm of large-scale on-site storage and remote processing of science data becomes feasible.

The paradigm is compelling because it focuses on enhanced return of information through the downlink of summary data products or analysis output, which can guide a more informed scheduling of data downlink. That is, by creating useful interim information products such as summaries or figures of merit on sensory data *en masse*, any existing downlink constraints can be better utilized by mission designers or scientists. Thus, the Nebulae paradigm is complementary to any improvements in communication technologies and can potentially free instrument designers from stringent data downlink constraints.

## 1.1 Workshop Summary

The Keck Institute for Space Studies hosted two workshops, with a total of 39 participants. This report represents the findings of those two workshops, and provides justification for this complementary operations paradigm. We provide concept descriptions, use cases, and pragmatic considerations. But first, we summarize the state of the art.

[1] Regrettably, this term is quite overloaded. JPL itself has at least one other "NeBula" initiative concurrent with this study.

# 2. State of the Art

Today's space missions are designed to return the maximally feasible amount of useful scientific information with high reliability at a minimal cost. Missions are nearly always funded and designed in a silo: their hardware, software, and scientific instruments are developed for a single mission with a defined duration, and self-contained objectives and resources. This approach is justifiable from a cost and risk management perspective, and scientists are able to use the collected data to pursue their particular questions. However, in order to achieve these lean and risk-managed architectures, trades are often made that reduce the amount or quality of returned data.

The primary reasons for these trades are based on physical limitations: light-time delay between the spacecraft and Earth, bandwidth for data transfer, and limited capacity for onboard data processing. The light-time delay prevents mission operators on Earth from making near-real-time decisions for the mission. In deep space, where the latency can be significantly longer than near Earth, missions must be designed such that human decisions are not time-critical. This has resulted in a *synchronous* or **transaction-based** command model for deep space missions to date: the operator on Earth issues a command, the spacecraft executes the command and returns the result, the operator reviews and analyzes, and the process repeats. This results in very slow data acquisition and long mission durations, and requires human operators to be in the loop, albeit light-time-delayed, at all times.

Data transfer bandwidth is generally a function of distance from Earth: as the spacecraft's distance doubles, the bandwidth quarters (assuming all else being equal). Therefore, deep space missions, particularly planetary and interstellar missions, are extremely constrained inthe amount of data they can send back to Earth. Mission designers are therefore tasked

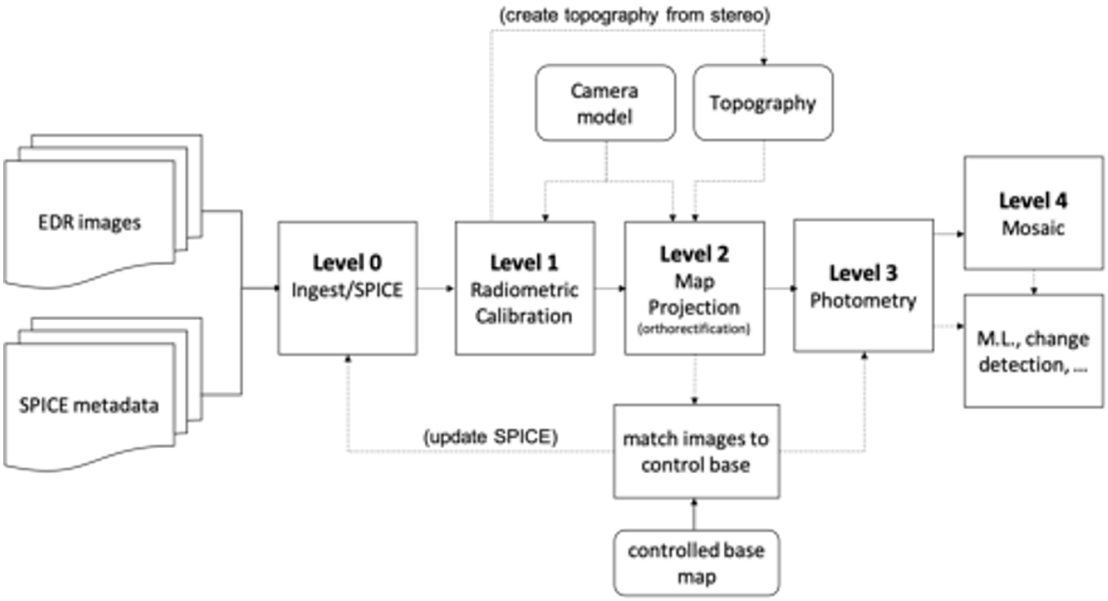



Figure 2.1: The ISIS3[a] pipeline as an example of science data processing. Very generally speaking, only the image/sensor capture (far left) occurs onboard the spacecraft currently. The workshop's main recommendation is to progressively move more of Level 0 through evel 3 processing on board as a "service" for science instruments, allowing higher-level data products to serve as summaries of the data that cannot be downlinked. These higher "summary" products are to be downlinked alongside the raw data sent to Earth.

[a]K. L. Edmundson, D. Cook, O. Thomas, B. Archinal, and R. Kirk, "Jigsaw: The isis3 bundle adjustment for extraterrestrial photogrammetry," *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. 1, no. 4, pp. 203–208, 2012.

with determining how to return only the most valuable data, which often results in the loss of "less valuable" data. As a greatly simplified example: imaging decisions for the High Resolution Imaging Experiment (HiRISE) camera onboard the Mars Reconnaissance Orbiter (MRO) start with a review of lower-resolution images from the context camera (CTX). The human operator reviews and requests high-resolution images of a subset of regions, and MRO returns only those.

Clearly, the total amount of data that makes it through this prioritization filter and back to Earth is much less than the amount of data that could be observed by the onboard instruments. For this reason, instruments are often designed for an artificially low duty cycle. Essentially, the sensing "capture rate" (the amount of data captured on an average operational day) will shrink to accommodate a lower downlink rate. Continuing the example, during its lifetime, MRO has returned 4% of the Martian surface in high resolution (at the time of this writing), despite years of flying over every part of Mars.

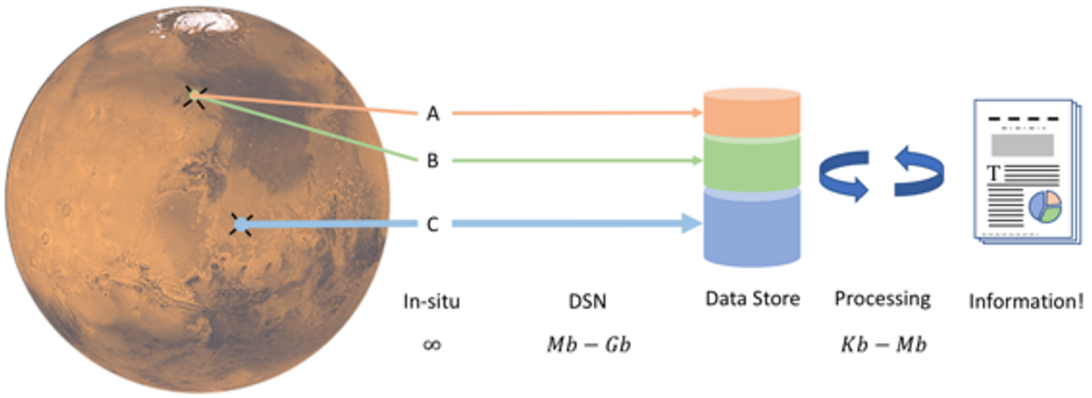



Figure 2.2: Under current paradigms, information return is limited by the data downlink rate set by the Deep Space Network, but at a rate much less than the instruments can collect. A select set of raw or lightly processed data products are returned to Earth, where high density information is then extracted and synthesized for publication by the mission science team.

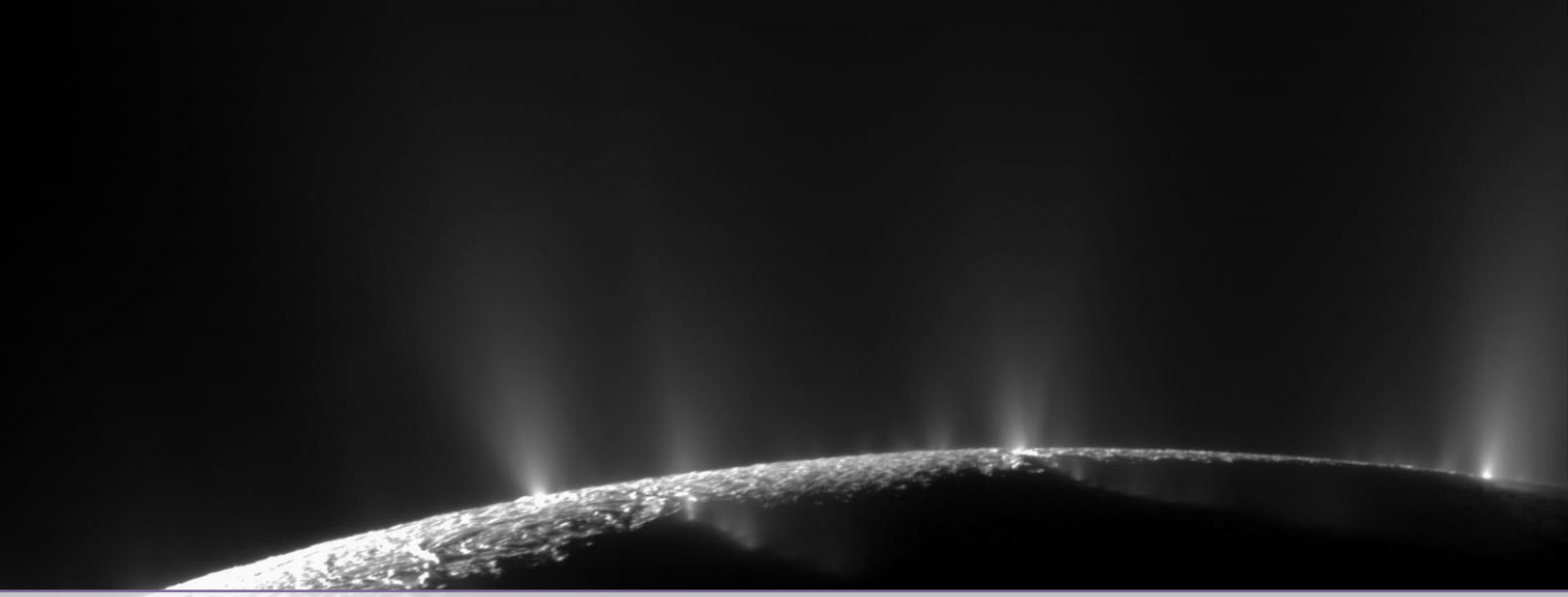
Note, data transfer bandwidth is perceived to be much less of a constraint for near-Earth missions, and because these missions have fewer perceived constraints, they are designed to produce an enormous amount of data. However, in the next chapter we revisit these assumptions.

In no mission concept could we find that the sensing "data rate" was not scaled to nearly match the downlink data rate, amortized over the course of a mission. That is, with only minor exceptions for image compression or pre-processing to allow prioritization, all data captured was intended for downlink.

After downlink to Earth, processing the data can also be a bottleneck on the way to producing useful science. For example, an Earth science satellite in low-Earth orbit (LEO) might be able to send gigabytes of image data every day (from the Terabytes of raw, uncompressed images it could take), but identifying characteristics of specific images that inform a study may take days of processing. This final step can take many reinterpretations, re-renderings, recombinations, or redistributions of the raw data that was downlinked, all typically done by iterating with data stored in a cloud server connected to a network of institutions, such as can be found with the Planetary Data System.

Whenever possible, the information gained from this last iterative step is fed back into the mission planning activities, to determine what data should be captured and/or downlinked next. Significant delays are potentially introduced if the spacecraft must reposition to satisfy the next identified need, or if a long queue of requests are ahead of this newly discovered need. This "loop closure" delay can be months-long.

In summary, the coarse mental model we propose for uncrewed missions to date is that of a remotely operated instrument which streams data back to a server for intensive processing before being re-tasked with another targeted observation. In this model, the two most important parts of the spacecraft are the instrument and the communications subsystem, because improvements to either of these two items have a direct impact on science return.



# 3. The Nebulae Mental Model

The main contribution of the workshop was to gather supporting evidence and motivating use cases for an expanded mental model of doing science using spacecraft. In the Nebulae mental model, missions aim to make use of data that would otherwise have been lost to architectural trades *and* to alleviate the loop-closure delay by enhancing onboard computing, increasing data storage, and prioritizing *information* transfer. This can begin humbly and alongside traditional operations, and can scale up as trust is built. Recognizing the need to reduce the risk to early adopters of Nebulae-like operations concepts, the intention is to begin with a minimal viable product (MVP) and allow the technology to mature over time. In this section, we outline one possible progression. The team captured variations on the Nebulae concept, ranging from single-instrument, to single-platform, to multiple-platform manifestations, and how these would stretch and impact the support needed from Nebulae capability. The team also examined architecture constructs that flow from the Nebulae concept. Lastly, the team considered how certain relevant and necessary architectural properties could be achieved.

The team examined use cases for the concepts associated with several NASA missions: MRO, Kepler, the Orbiting Carbon Observatory 2 (OCO-2), and several others. Finally, the team looked further at the special case of the future of observing Earth as a System. As noted, this capability is evolving, moving toward continuous spatial and temporal coverage, with space platforms, sensors and instruments, data provisioning, and possibly elements of the Nebulae concept (e.g., flight computing) all becoming commoditized. With bandwidth for ground communications generally being ample (although still subject to race conditions against data collection capacity), what are the drivers for adopting Nebulae in an Earth-observing context? The team examined this question, along with the corollary of seeking the most effective

partnership between government and industry, leveraging the strengths of each and with awareness of the divergent objectives of each.

Specifically, we introduce four computing-intense mission concepts: 1) Vigilant Instrument; 2) Data Server in the Sky (with the technical challenge to make such a server as reliable and trusted as any ground-based data archive); 3) Observing System in the Sky, examining anticipated continuous spatial and temporal observing coverage of Earth as a System, and whether the capability lends itself to users becoming subscribers, as a departure from the traditional uplink/downlink transaction model for interacting with space-based systems; and 4) the Data Cycler (or Data Mule), an adjunct concept meant to address concerns about returning to the ground remotely held data in a timely fashion.

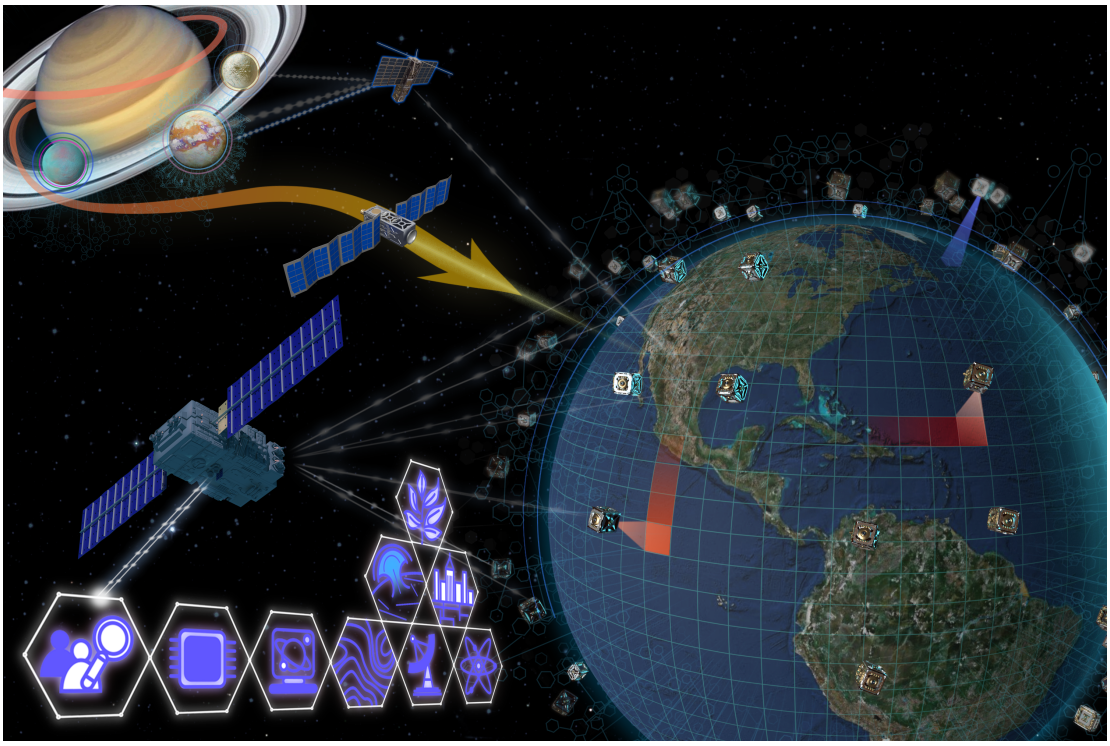


Figure 3.1: The mental model of a spacecraft changes from a remote instrument, or "transactional" model (left), to an in-situ repository of significant data that can be interacted with (right), being constantly updated with in-situ observations.[a]

[a]Vander Hook, J., Castillo-Rogez, J., Doyle, R., Vaquero, T. S., Hare, T. M., Kirk, R. L., ... & Cocoros, A. (2020, March). Nebulae: A Proposed Concept of Operation for Deep Space Computing Clouds. In *2020 IEEE Aerospace Conference* (pp. 1-14). IEEE.

## 3.1 Vigilant Instruments

A **Vigilant Instrument** refers to a concept of a sub-system that can produce in-stream summaries of data that is otherwise not requested for immediate downlink. In its simplest

form, it is a relatively powerful computing resource that can be paired with an onboard instrument in order to provide additional data processing or storage. An example use case would be a camera that can process images in real time to report the presence of features of interest such as craters. Such an instrument would provide two data products: A high volume series of images on the order of gigabytes/second, and a lightweight "summary" of those images which may be three or more orders of magnitude more compact. The summary informs which images contained the particular feature of interest and the location in the image of the feature. The lightweight product can increase the value of the main image sequence by informing downlink priorities, triggering capture of full-size or thumbnail images for a limited-size cache of "bonus" downlink, or simply by reporting the location of suspected features as part of source collecting/processing.

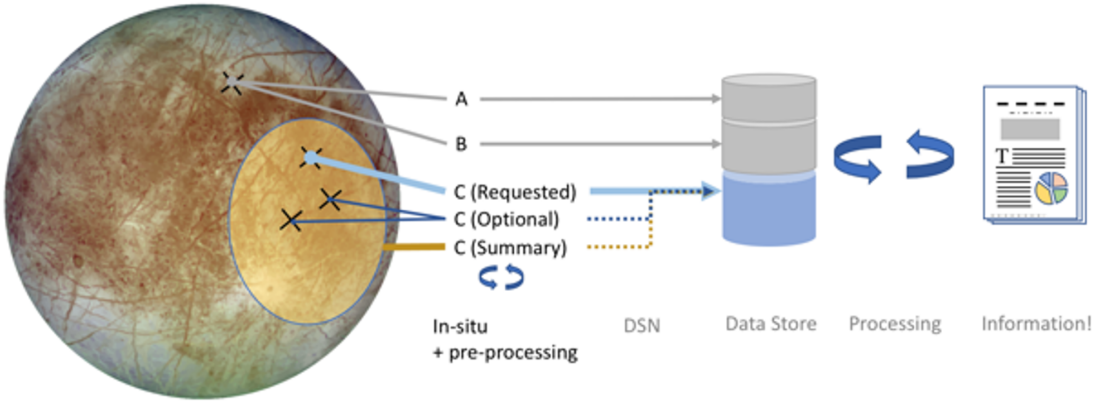



Figure 3.2: Increasing the computational power on board an asset increases the quantity and quality of information returned per bit through the Deep Space Network. In-situ processing allows selection from a set of possible measurements using figures of merit ("take this image if there are new craters visible"), or even a summary of an entire area of a planet using those figures of merit ("how many craters are here and where are they?"). This enables more rapid identification of high-priority targets and observations, both increasing the possible scientific return and reducing the latency in processing and analysis by the science team.

### 3.1.1 Vigilant Instrument Retrospective: Kepler

The Kepler mission was launched in 2009 and was a large space telescope designed to search for Earth-sized planets orbiting other stars. It did so using a single photometer that constantly monitored the brightness of approximately 150,000 stars within a fixed field of view. The onboard instrument measured the brightness of the stars (and only the stars) once every thirty minutes. The brightness sequences were transmitted to Earth, where they were processed to determine the probability of a "dip" corresponding to a planet transiting the star and blocking some of the light.

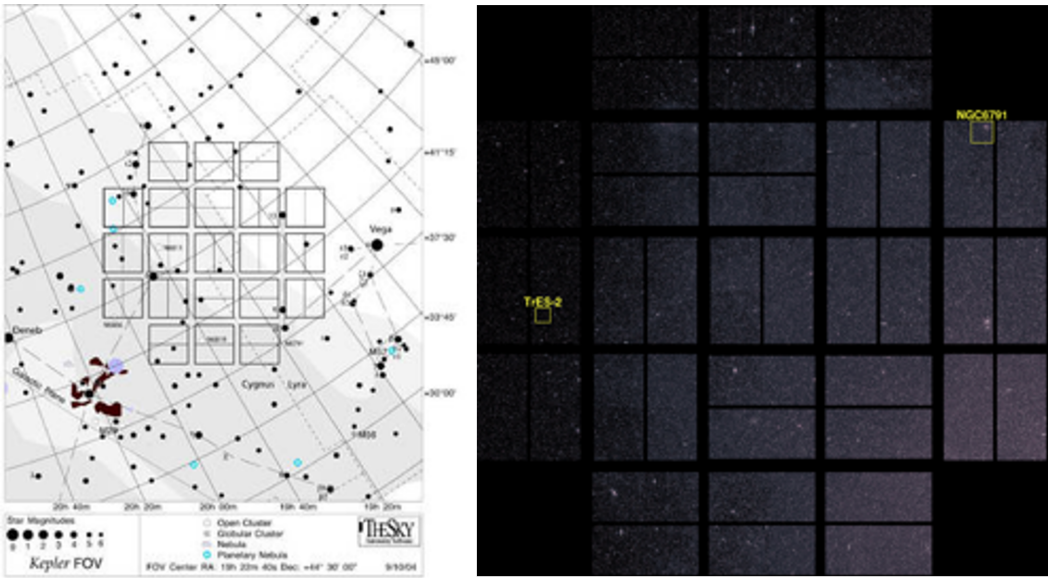



Figure 3.3: Field of view of the Kepler space telescope. CREDIT: NASA/Ames/JPL-Caltech

In a sense, this is a Vigilant Instrument. It is constantly observing to catch scientific data that depend on temporal variation, and it does so for many targets simultaneously. A retrospective redesign of Kepler including more compute and memory resources allows us to expand its science goals. For example, given onboard compute power and storage that is sufficient to evaluate many sequential images of the whole focal plane, not just the brightness of the stars, we could use synthetic tracking to search for moving objects as well.

Synthetic tracking is a computational technique that shifts and adds a sequence of images to find moving objects.[1] Given Kepler's wide field of view (115 deg$^2$), it would provide a valuable resource for surveying near-Earth asteroid activity. The importance of this type of survey has been emphasized in recent decadal surveys due to the danger of a large, catastrophic collision.[2].

## 3.2 Data System in the Sky

A **Data System in the Sky** refers to the concept of a possibly remote and highly reliable science data archive, which forms a crucial element of a Nebulae deployment, including in the proximal environment. The archive serves not only to meet the needs for long-term data storage, but may also support ongoing application of analytics in a remote environment, leveraging already-captured data as a baseline. Such an archive can also support mission

[1] Zhai, C., Shao, M., Nemati, B., Werne, T., Zhou, H., Turyshev, S. G., ... & Harding, L. K. (2014). Detection of a faint fast-moving near-earth asteroid using the synthetic tracking technique. *The Astrophysical Journal*, 792(1), 60.

[2] Britt, Daniel, et al. "Community White Paper to the Planetary Science Decadal Survey, 2011–2020."

concept evolution in a remote environment without the need to deploy additional space platforms. Data Systems in the Sky must survive for long periods and preserve functionality without degradation in order to be useful.

### 3.2.1 Data System in the Sky Retrospective: Mars Reconnaissance Orbiter

The Mars Reconnaissance Orbiter was launched in 2005 and began its main mission from Mars orbit in 2006. Among many instruments, it carried with it two visible spectrum imagers, the Context Camera (CTX) and the High Resolution Imaging Science Experiment (HiRISE), as well as the Compact Reconnaissance Imaging Spectrometer for Mars (CRISM). From 300–400 km altitude, CTX could capture Mars' terrain at about 6 meters/pixel, and HiRISE at ~1 m/pixel. CRISM was capable of imaging long swaths of terrain in a low-resolution mode (100 m/pixel) or focused areas at higher resolution (10 m/pixel). For the purposes of our retrospective, we limited ourselves to an examination of the imagers.

By all accounts, MRO has been a resounding success. It has gathered the spectral, topographical, and imagery data that has enabled revolutionary discoveries about the history and current-day processes of Mars. Over the lifetime of MRO's primary mission and extended operations, it has returned a total imagery data volume on the order of 300 Tb. The amount of data returned corresponds to coverage of the planet as shown in Table **??**, tallied as of the time of this writing. By far, the dominant use of imagery downlink capacity has been with HiRISE.

Notably, this total downlink budget for the main imagers would now fit on one or two modern solid-state drives, weighing less than 100 grams. The reimagined MRO main mission that makes use of the Data System in the Sky concept would use, perhaps, 10$\times$ this amount of onboard storage to collect "extra" imagery. The ancillary imagery would be used to perform onboard change detection, downlinked as thumbnails, or used to efficiently gather statistics about the changes of the whole of the Martian surface over time.

To reinforce this point, consider one of the most compelling mysteries disclosed by the main MRO mission, known as Recurring Slope Linae (RSL). These periodic dark streaks were visible on Mars seasonally and are thought by some to be signs of flowing water. The decades-long investigation of HiRISE imagery for RSL involved looking for changes over time in subsequent imagery taken days or even weeks apart. One could envision a system deployed to Mars that was equipped with an "always-on" imager (perhaps with the form factor of CTX), and onboard storage sufficient to keep a deep history of the observations of a large swath of the planet. In the event of a serendipitous discovery of a phenomenon of interest, the system could be tasked with returning the data from its storage and summarizing a history of the changes from the region. It could also do rough parameterized matching over a much larger part of the planet, all onboard, returning these insights well before the satellite travelled back

to the area to collect follow-up imagery. This could have accelerated our decades-long study of RSLs significantly by allowing near-instantaneous interrogation of changing areas of Mars.

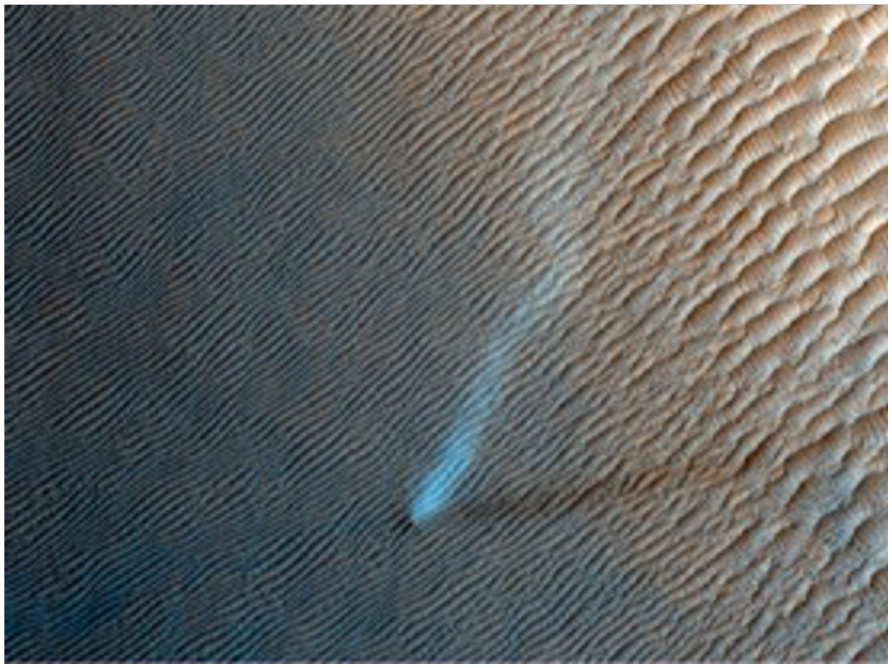


**Figure 3.4:** Dust devil on Mars. Credit: NASA/JPL-Caltech/Univ. of Arizona

Carried to its extreme, 2 Pb would be sufficient to hold in memory a raw image of the entire surface of Mars, or about 1/10 that, given the JPEG compression scheme used. Studies are already underway in deploying automated crater detection to Earth-side Mars imagery.[3] Deploying such technology to the next Mars orbiter, coupled with on-site processing, would be a perfect example of the Vigilant Instrument concept. What the Data System in the Sky provides is history, context, and robustness to discoveries or changing priorities through the use of a large volume of extra measurements held in reserve.

The inclusion of computing and onboard storage exclusively to generate additional useful data products or cache imagery could have had a significant impact on MRO's discovery timeline, without impacting the traditional investigation methodologies, including manual targeting and iterative discovery. The end result might have been a likely acceleration of discovery by the same scientists, using the same methods on the same data, but simply with access to more of it, more quickly, and with bonus data products assisting their investigations.

[3]The COSMIC team (`https://ml.jpl.nasa.gov/projects/cosmic/cosmic.html`) recently published a press release on their work finding fresh craters on Mars: `https://www.jpl.nasa.gov/news/ai-is-helping-scientists-discover-fresh-craters-on-mars`

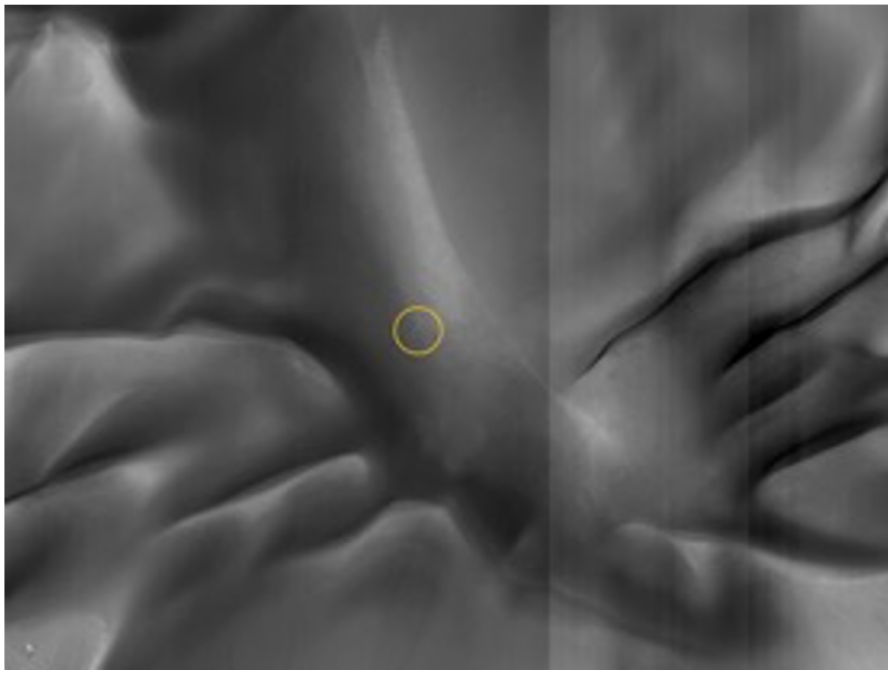


Figure 3.5: Wide view of the location of the Martian dust devil shown in Figure 3.4.

## 3.3 Observing System in the Sky

The **Observing System in the Sky** concept emerges from trends in studying Earth as a System, and is enabled by the increasing commoditization by industry of space platforms, sensors and instruments, data provisioning, and even Nebulae-relevant resources such as computing. The key advance is moving to nearly continuous spatial and temporal coverage—often through the use of multi-platform networks of sensors. Architecturally, these trends can allow a shift whereby users (scientists, decision makers, individual business owners such as farmers, etc.) become direct subscribers of the observing service, distinct from the traditional uplink/downlink transaction-based mode of interacting with space platforms. The concept can in principle be extended to remote environments via Nebulae, with "users" leveraging deployed space assets endowed with various autonomy and analysis capabilities. Because the technology is expected to change as Nebulae matures, it is vital that early designs follow architectural patterns that allow Nebula-enabled spacecraft to remain operational, and interoperate with newer spacecraft for as long as possible, in order for Nebulae to properly scale.

Enhanced computational capability onboard spacecraft would significantly increase the quantity of information per bit that can be returned to Earth by enabling data fusion and information summarization beyond simple compression. Improved capabilities include (a) the storage of more raw and summarized data for downlink or eventual full transfer via a "data-return" mission; (b) science goals that depend on such data fusion capabilities; (c) calibration, spatial alignment, and enhancements to newly acquired datasets using pre-loaded older lower-

resolution datasets; (d) greater autonomy in tasks like aligning new images to an existing base map; and (e) more diverse summary products based on change detection and longer temporal baselines.[4]

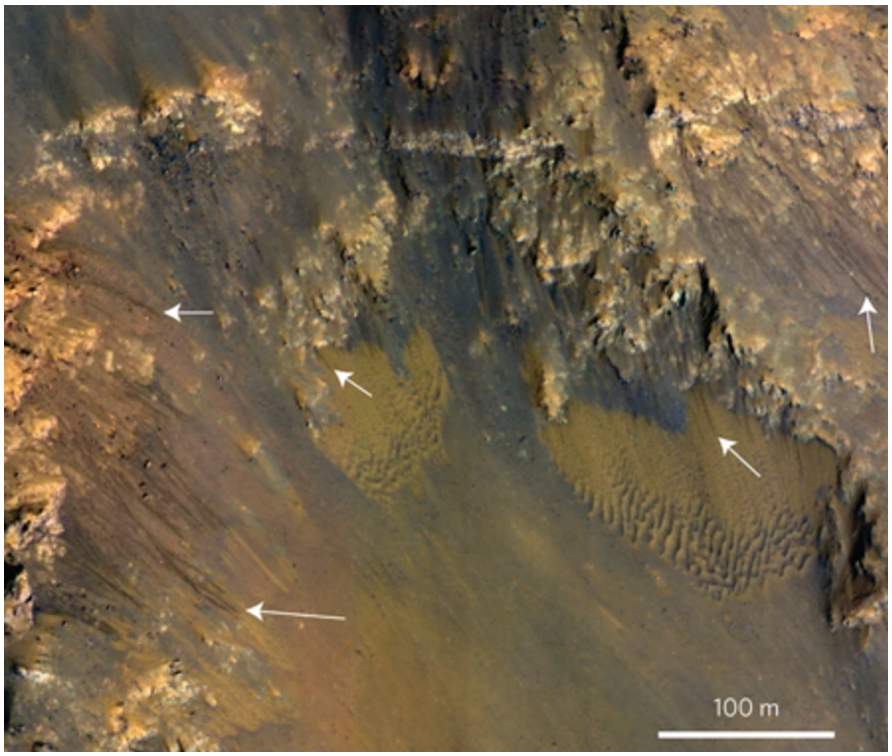



**Figure 3.6:** Annotated HiRise image showing RSLs as dark streaks. Credit: NASA/JPL/UArizona, Wikimedia Commons.

[4]Mahabal, A., Hare, T., Fox, V., & Hallinan, G. (2021). In-space Data Fusion for More Productive Missions. *Bulletin of the American Astronomical Society*, 53(4), 500.

| Reference | Coverage | Resolution | Bands | Volume | Cadence | Data/time (Kbps) |
|---|---|---|---|---|---|---|
| HiRISE [1] | 4% | ~1 m/pixel | 3 | 268 Tb | 15 years | 556 |
| CTX [2] | 99% | ~6 m/pixel | 1 | 62 Tb | 15 years | 131 |
| CRISM (targeted) [3] | ~1% | 15 m/pixel | ~100 | 7 Tb | 15 years | 14 |
| CRISM (untargeted) [3] | ~100% | 100 m/pixel | <50 | 10 Tb | 15 years | 21 |
| **Proposed** | **Coverage** | **Resolution** | **Bands** | **Volume** | **Cadence** | **Data/time (Kbps)** |
| Imaging Spectrometer [4] | 100% | 30 m/pixel | 200 | 2,000 Tb | 4×/year | >253,000 |
| Visible Imager [1] ] | 100% | 1 m/pixel | 3 | 6,700 Tb | 1×/year | >212,000 |

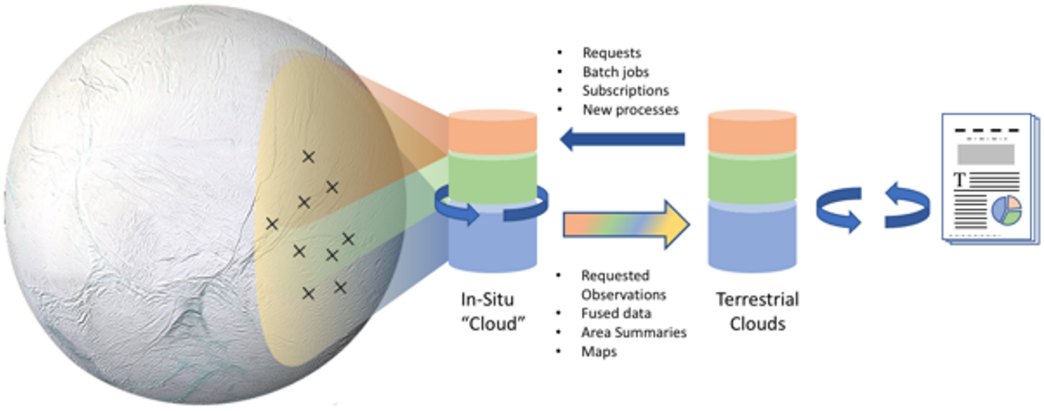



Figure 3.7: A fully realized Nebulae concept enables multiple analysis queries, alerts, or processes to exchange information between spacecraft, instruments, and terrestrial systems, on demand or driven by events, through the Deep Space Network. The operations concept maximizes the information potential of the available downlink. An in-situ "cloud" can be co-located with a system, or serve many systems.

## 3.4 Bonus Concept: Data Cycler

A **Data Cycler** (or Mule) is a shuttling orbiter between Earth and a remote environment for the express purpose of transferring data archived at the remote environment back to Earth for ease and timeliness of traditional ground-based investigation. The concept is similar to cargo shuttles conceived for the deployment of modules and materiel to, e.g., the Moon, and must involve a stable, cycling trajectory. The basic idea is to enable a high-bandwidth burst

of data when the Data Cycler is near the remote archive, and then efficiently move that data back into the Earth environment through the use of a low-thrust orbit that trades a higher transit time for a reduced requirement for onboard fuel, power, and/or propulsion.[5,6]

While the concept may seem far-fetched, the currently planned Mars sample return (MSR) campaign is a reasonable conceptual grounding. MSR is planned to include three large-scale missions: a sample cache rover (Perseverance) to collect soil samples from selected locations; a sample fetch rover to retrieve and load the samples into a lift vehicle which will return the samples to Mars orbit; and a final mission to rendezvous with the orbiting sample canisters, retrieve them, and return them to Earth. One can envision a much less complex mission consisting of a satellite launched onto a cycler orbit which allows low-thrust periodic revisits within Earth and Mars proximity. Since the Data Cycler's only requirement is high-data rate communication for short periods of time (i.e., when in proximity of Mars and Earth), and the rest of its life is spent safeguarding data through regular refreshes and error corrections, the cost of a Data Mule could be substantially reduced—even to the point of being a secondary payload. Given that the yearly data volume for even optimistic projects of the Deep Space Network is on the order of a commercial hard disk, a well-timed Data Mule could double or even triple the yearly return of the DSN with minimal impact to its operations.

The Data Cycler concept can be an adjunct to the Data System in the Sky concept to address possible concerns about the long-term viability of remote data. The general and agreed-on objective of ultimately returning all collected data to the ground is addressed by Data Cyclers. It is also important to note that, in the Nebulae sense, the remote archive is enabling science objectives to be pursued in the remote environment in ongoing fashion, with or without the eventual return of the remote data.

## 3.5 Discussion—The Faces of Nebulae

For these concepts, we have discussed benefits that could have accrued to historical and/or current NASA science missions if scalable and ample space-based computing, data storage, and networking capabilities, services and infrastructure—e.g., the Nebulae vision—had been available to those missions.

As described above, the basic objective of this activity was to pose the retrospective question: "What additional value-added science might historical and/or current space missions have accomplished if Nebulae-style capability had been available to those missions?" This exercise was conducted as a thought experiment across several NASA mission use cases, noting both

[5]Solar System Data Mules: Analysis for Mars and Jupiter. Marc Sanchez-net, Etienne Pellegrini, Wilson Parker, Joshua Vander Hook, *Proceedings of the IEEE Aerospace Conference*. Big Sky, MT 2021.

[6]Data Mules on Cycler Orbits for High-Latency, Planetary-Scale Data Transfers. Marc Sanchez-net, Etienne Pellegrini, Joshua Vander Hook, *Proceedings of the IEEE Aerospace Conference*. Big Sky, MT 2020.

the additional science prospects and the attendant Nebulae resources and configurations that would have enabled that science.

The thought experiments need not be an end in themselves, however conceptually valuable they may be. Asking these "What-If?" questions can lead to the formulation of technology experiments in which the discussed scenarios can be explored further. Specifically, any mission with high-fidelity testbeds and high-fidelity models (including simulation capabilities) of the space platform, instruments, and operating environment can pose well-formed questions about what could have been accomplished with Nebulae-style capabilities. Such experiments could extend to more or different forms of autonomy on the flight side with onboard analytics, along with the computing, storage, and network-related enhancements of Nebulae. Furthermore, these experiments can be further informed from detailed project archives with records of commands issued and telemetry received, how operations concepts evolved over the course of the mission, and how ground and flight capabilities evolved via software upgrades and/or hardware degradations. Historical and operating Mars surface missions have such detailed records and archives of this character and resolution.

Conducting such well-formed technology experiments in upcoming years can in turn promote the final objective of the Nebulae workshop series, namely the development of a mission-level capability proposal to demonstrate the Nebulae capability, ideally as an augmentation to existing deployed assets and resources.

## 3.6 Detailed Case Study of an Observing System Example: Earth as a System

From a retrospective viewpoint, observing Earth presents a unique opportunity. Continuous observations over several decades provide a plethora of real-time data streams, as well as archived datasets from which hypothetical Nebulae concepts—such as advanced data selection or compression algorithms—could be explored. Unlike deep space missions, Earth-observing and Earth-orbiting missions need not treat each science opportunity as precious, as would be the case with flyby or rover missions where there may only be a single chance to explore a specific location or subset of a distant planet's surface. Ample observations and a relatively continuous data record at Earth provide multiple observations of science phenomena. Nor are Earth-observing and Earth-orbiting missions severely downlink-constrained, as is the case with planetary or deep space missions. For the most part, Earth-observing and Earth-orbiting missions are able to downlink a large fraction of the data collected, with mission requirements dictating the amount of science data and telemetry that can be downlinked with sensor output sized accordingly. For this reason, retrospective analyses of Earth Science missions cannot address the missed science value of data that could not be downlinked or science collection opportunities that were missed due to mission design. Instead, retrospective analyses can address the scientific value of multiple measurements of a science phenomenon observed from different viewpoints/orbits and via different sensor modalities. A perfect working example of

this trend is encapsulated in the "Sensor Network" concepts pioneered by the AI group at JPL.[7]

### 3.6.1 Nebulae concepts for Earth

Mission requirements have traditionally dictated the amount of data that can be downlinked and sensor data output has been sized accordingly. As a consequence, some onboard processing or scientist-in-the-loop (SITL) data selection is typically done if the amount of science data produced onboard is substantially greater than the downlink bandwidth can accommodate. Examples of this include the aforementioned OCO-2 mission as well as the Magnetospheric Multiscale (MMS) mission. With the MMS mission, both a high-rate (i.e., "burst" mode) and low-rate (i.e., "survey" mode) data stream are collected, with mission design allowing all of the survey data but only 4% of the high-rate data to be downlinked.[8] Various SITLs examine the survey data and prioritize time intervals of burst-mode data that are desired to be downlinked, with a requirement that selections must be made within 12 hours of observation time. The MMS spacecraft are only capable of holding 48 hours of burst mode data in memory, with the burst mode data continually being overwritten. Therefore, scientists must be continually in the loop and manually inspecting data to ensure the most relevant observations are retained and downlinked. Only recently has AI been explored as a means for automating some of the science data selection currently done manually by SITLs, with the intent to alleviate the time needed for scientists to make data selections. Similar to the case of onboard processing and compression with OCO-2, the inability to get all of the high-rate data captured by the MMS mission to the ground means that scientifically relevant data might be lost.

Yet another example is the Thermosphere Ionosphere Mesosphere Energetics and Dynamics (TIMED) mission, specifically the Global Ultraviolet Imager (GUVI). Launched in 2001 (and subject to the hardware limitations of the time), GUVI was capable of collecting data over 14 spatial pixels and 160 spectral bins ranging from 115 to 180 nm. However, due to downlink constraints, only a small subset of the spectral data was able to be downlinked.[9] Thus, while it is generally assumed on the basis of proximity that Earth orbiting missions are able to downlink all of the data captured, we find that these missions indeed face downlink limitations that have impacted the science return and Concept of Operations (CONOPS) design. Both OCO-2 and MMS have been designed to cope with such limitations by employing Nebulae-like

[7] Chien, S. A.; Davies, A. G.; Doubleday, J.; Tran, D. Q.; Mclaren, D.; Chi, W.; and Maillard, A. Automated Volcano Monitoring Using Multiple Space and Ground Sensors. *Journal of Aerospace Information Systems (JAIS)*, 17:4: 214-228. 2020.

[8] Argall, M. R., et al. MMS SITL Ground Loop: Automating the Burst Data Selection Process. *Frontiers in Astronomy and Space Science*. 7:54. 2020.

[9] Paxton, L. J. et al. Global Ultraviolet Imager (GUVI): Measuring the Composition and Energy Inputs for the NASA Thermosphere and Ionosphere Mesosphere Energetics and Dynamics (TIMED) Mission. *Proc. SPIE., 3756, Optical Spectroscopic Techniques and Instrumentation for Atmospheric and Space Research III*. 1999.

concepts—OCO-2 by employing compression techniques onboard and MMS by implementing a "Data System in the Sky" concept.

As Earth-observing missions move toward larger constellations and data-intensive sensors, incorporation of Nebulae-like constructs on board space platforms or within a distributed ground system will become ever more critical. While downlink constraints have largely not been a driving factor in Earth-orbiting missions as of yet, when it has been a factor, missions have had to carefully consider how to down-select or compress sensor output to get the most relevant information to the ground. The proposed next generation of sensors, such as hyperspectral sensors and synthetic aperture radar (SAR) imagery, will produce vastly larger datasets over short operational timelines. Current proposed hyperspectral CubeSat missions, such as the Compact Hyperspectral Air Pollution Sensor-Demonstrator (CHAPS-D), are only able to downlink a few minutes of uncompressed sensor data per day of flight time.[10] With such sensors operating on a singular spacecraft or multiple spacecraft, CONOPS will need to be considered that maximize downlink of the most relevant science information—this could manifest in targeted observations, onboard compression algorithms, or downlink of derived products only ("Data System in the Sky" concept again). As one such example currently underway, industry CubeSat constellations that are producing vast quantities of EO/IR data have designed their CONOPS to only image over land masses, using intervals over the ocean to duty-cycle and downlink observations.

Ground stations of the near future will need to incorporate Nebulae-like concepts to handle the vast amount of data flowing in from large CubeSat constellations and data-intensive sensors. These ground stations will need to prioritize assimilation of data based on data quality and scientific/mission value as well as produce fused or derived data products leveraging observations from multiple missions. Numerical Weather Prediction (NWP) is one such example of a fused data product, requiring assimilation of relevant data products with strict timeliness requirements to produce valid and critical NWP model runs. NASA GMAO runs an operational NWP product and has suffered in the past from assimilating data of poor quality, resulting in forecasts with substantial errors.[11] Pre-screening vast quantities of data with strict timeliness requirements while maintaining data quality is already a recognized need in the community, both for NASA GMAO as well as NOAA.[12] With future Nebulae concepts, such screening algorithms could be present not only on the ground but on board the spacecraft as well. One such example of data screening done on board could be autonomously recognizing an image has substantial cloud cover and consequently little scientific value. By

[10] Swartz, W. H., et al., CHAPS: A New Compact Instrument for Air Pollution Remote Sensing. Fourth Conf. on Earth Observing SmallSats, *Amer. Meteor. Soc. Meeting*. January 13, 2021.

[11] NASA GMAO (2016). Erroneous CO emissions over California cause unrealistic CO concentration in GEOS-5 model. *GEOS System News*. March 1, 2016.

[12] US Dept. of Commerce, NOAA. NOAA Data Strategy: Maximizing the Value of NOAA Data. July 2020. `https://nrc.noaa.gov/Portals/0/2020%20Data%20Strategy.pdf?ver=2020-09-17-150024-997`

autonomously recognizing that a subset of data does not fulfill science objectives (even in the simplest realization previously mentioned), spacecraft could prioritize downlink bandwidth to the most relevant measurements.

Timely event detection could allow for "tip and cue" constellation management, in which routine observations cue a more capable sensor or different sensor modality to investigate an event further, or simply cue additional observations from the same satellite. Timeliness requirements would dictate the degree of automation required, with the need for more rapid response from disaggregated constellations requiring a fully-enabled Nebulae configuration. To accommodate such strict timeliness requirements, Nebulae-like ground systems and constellations will need to automate detection of anomalous events and subsequent coordination of constellation assets. Automation may initially be present in the ground systems due to the ease of deployment, but as the TRL increases and cross-linked constellations are launched, it is more effective that such automation will take place onboard the spacecraft. Examples of such "tip and cue" constellations can already be found in industry with SITL involvement in constellation management. In these examples, the cueing of follow-on observations is focused on capturing data from different and more capable sensors rather than meeting strict timeliness requirements to capture an evolving phenomenon.

#### 3.6.1.1 Differences in industry versus NASA/NOAA approach in Earth-observing/orbiting constellations.

The future of Earth-observing missions is in higher spatial and temporal coverage, fused data products from multiple types of sensor modalities, and more advanced sensors, such as hyperspectral imagers, capable of producing vast data volumes. Industry-led LEO CubeSat constellations have dramatically increased in size in the past decade, providing the increased temporal and spatial sampling cadence desired for many science applications. For example, Planet has a constellation of more than 150 satellites that capture all landmasses on Earth daily with 3.7-m resolution and include taskable assets able to obtain imagery at a 50-cm resolution anywhere on Earth twice daily.[13] Commercial sensor payloads have primarily been imaging systems in RGB and NIR as well as Global Navigation Satellite System Radio Occultation (GNSS-RO). Given the low cost of CubeSat constellation deployment, industry-led efforts have embraced "agile" technology development. Continuous, incremental improvements are made to CubeSat hardware, and new "tranches" or "flocks" are frequently launched. Failures are accepted as part of the innovation process and budgeted for, as is planned obsolescence. As a consequence, the latest technology is able to be flown and tested in a rapid development mode, allowing industry to continuously iterate and improve flight hardware and software as well as the supporting ground infrastructure. Current operations concepts for industry constellations have been designed to downlink all of the data collected, with duty cycles and downlink time budgeted for orbit locations that have little commercial value (i.e., remote

[13] Planet. Planet Imagery Product Specifications. February 2021. `https://www.planet.com/products/`

maritime locations). The ground infrastructure has been designed to accommodate vast quantities of data from multiple flocks, and produce "subscribed" derived products for end users. As a consequence of a business model focused on collecting and downlinking as much data as possible, with all analysis occurring on the ground, earth-sensing industry partners have not developed onboard processing technologies.

In contrast to industry, the Department of Defense (DoD) and NASA have remained focused on larger, more capable satellites and the development of new sensor technology. Large missions, such as those within the NASA A-Train (Aqua, Aura, OCO-2, and GCOM-W1) and the DoD's missile warning program have remained the main technology drivers. Onboard analytics have focused on compressing sensor output to the downlink constraints set at the mission level. As of yet, the level of compression needed has not required advanced artificial intelligence (AI)/machine learning (ML) techniques, but the next generation of sensors will undoubtedly challenge that precedent. While CubeSats have begun to be embraced within the government sphere, development efforts have been focused more on miniaturization of sensor technology, as seen with the NASA Instrument Incubator Program (IIP), rather than large-scale constellation deployment as seen in industry. The first large-scale government constellation won't be launched until 2022. As a consequence, NASA technology development has fallen behind industry with regards to developing ground infrastructure to support large-scale constellations, including not only daily operations but infrastructure to ingest, store, and perform analytics on vast quantities of data. Similarly, the inability to embrace failure as an option within NASA has led to slowed TRL pipelines for CubeSat missions and a delay in embedding data analytics and AI within ground systems or onboard spacecraft.

#### 3.6.1.2 Leveraging industry partnership to enhance Nebulae concept demonstration and development.

NASA's strengths lie in the development of advanced sensor technologies and miniaturization of technology for CubeSat platforms, as demonstrated via the IIP. In contrast, industry has paved the way towards rapid, scaled deployment of new CubeSat technologies and back-end ingestion and analysis of data products. Partnerships between NASA and industry to rapidly deploy not only sensors but potential "smart" algorithms within ground infrastructure or onboard, validated via technology demonstrations, would be of great value. Such partnerships would allow for NASA to represent the interests of the broader science community dovetailed with industry technology development efforts. This approach would ensure that the sensors deployed in large CubeSat constellations produce data of sufficient quality for science end-users and similarly that technology development would progress towards sensor modalities that are beneficial to science. Recent science efforts focused on merging USGS LandSat imagery with Planet imagery found there was not adequate calibration of the sensors on Planet's Cubesats either before launch or while on-orbit. This led to difficulties not only in comparing Planet imagery with LandSat imagery but also in comparing imagery from various

CubeSats in Planet's constellation.[14] Nevertheless, the science community has recognized the potential of Planet (or other commercial) imagery in bridging the gap between ground observations and current, lower-resolution imagery from existing NASA and NOAA satellites. By encouraging further collaboration between industry and science end-users, the quality and utility of commercial datasets will only be enhanced.

For software development efforts, NASA could view industry partnerships as a "testbed" environment, particularly for large-scale constellation efforts. Future Nebulae concepts consisting of fused data products from multiple missions, or constellation management techniques, can be explored via the industry's existing CubeSat constellation and ground system infrastructure. Industry has substantially more expertise in data processing pipelines and deployment of algorithms at scale, albeit once the data has already reached the ground. Earth Science application areas that can benefit from fused data products (i.e., fusing data from NASA missions and industry data products) or "tip and cue"-style CONOPS could be explored as a means for collaboration between industry and NASA. Such a collaboration could initially be explored in a retrospective sense to explore the scientific value of a fused data product, and subsequently deployed within industry ground systems as an ancillary data pipeline. One such example is Atmospheric Motion Vector (AMV) winds, which are derived from NOAA GEOS imagery and currently a stove-piped product. 3D atmospheric winds were highlighted by the 2018 Earth Science Decadal Survey as a highly desired and poorly measured phenomenon, with accurate, global 3D wind determination crucial for NWP.[15] Including imagery from additional NASA EO/IR sensors as well as the vast imagery datasets available from commercial entities could dramatically improve AMV determination, since AMV retrievals benefit from multiple look angles and rapid temporal and spatial updates. Commercial entities could provide the spatial and temporal coverage currently lacking in AMV detection. Initial retrospective studies could explore the improvement in AMV determination that could be possible with the inclusion of commercial imagery. Subsequent Nebulae concepts could be explored by producing a fused data product on the ground (leveraging industry expertise), porting AMV determination onboard the spacecraft, and "tip and cue" constellation management techniques.

While both NASA and industry have lagged in porting computation onboard the spacecraft due to hardware limitations and lack of downlink restriction, future Nebulae concepts will require onboard advanced computation. Earth-observing CubeSat missions should be viewed as a low-risk opportunity to test hardware and algorithms needed for eventual deep-space Nebulae

[14] Johnson, B. R.; McGlinchy, J.; Cattau, M.; Joseph, M.; and Scholl, V. Harnessing Commercial Satellite Technologies to Monitor Our Forests. Proceedings Volume 10767, Remote Sensing and Modeling of Ecosystems for Sustainability, *SPIE Optical Engineering and Applications*, San Diego (2018).

[15] National Academies of Sciences, Engineering, and Medicine. 2018. *Thriving on Our Changing Planet: A Decadal Strategy for Earth Observation from Space*. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/24938.

concepts. With NASA-developed instrument payloads on industry-supplied CubeSat buses, onboard analytics for compression and data screening could be embedded within the instrument payload and incremental tiers of capability/complexity tested for particular science use cases. Thus, CubeSat instrument development could be explored as a means towards improving the onboard compute capabilities and raising the TRL of various algorithm approaches. Embracing industry's agile development schedule through NASA-industry partnership could enable rapid TRL improvement while maintaining a managed risk posture.

#### 3.6.1.3 Nebulae concept development through NASA resources

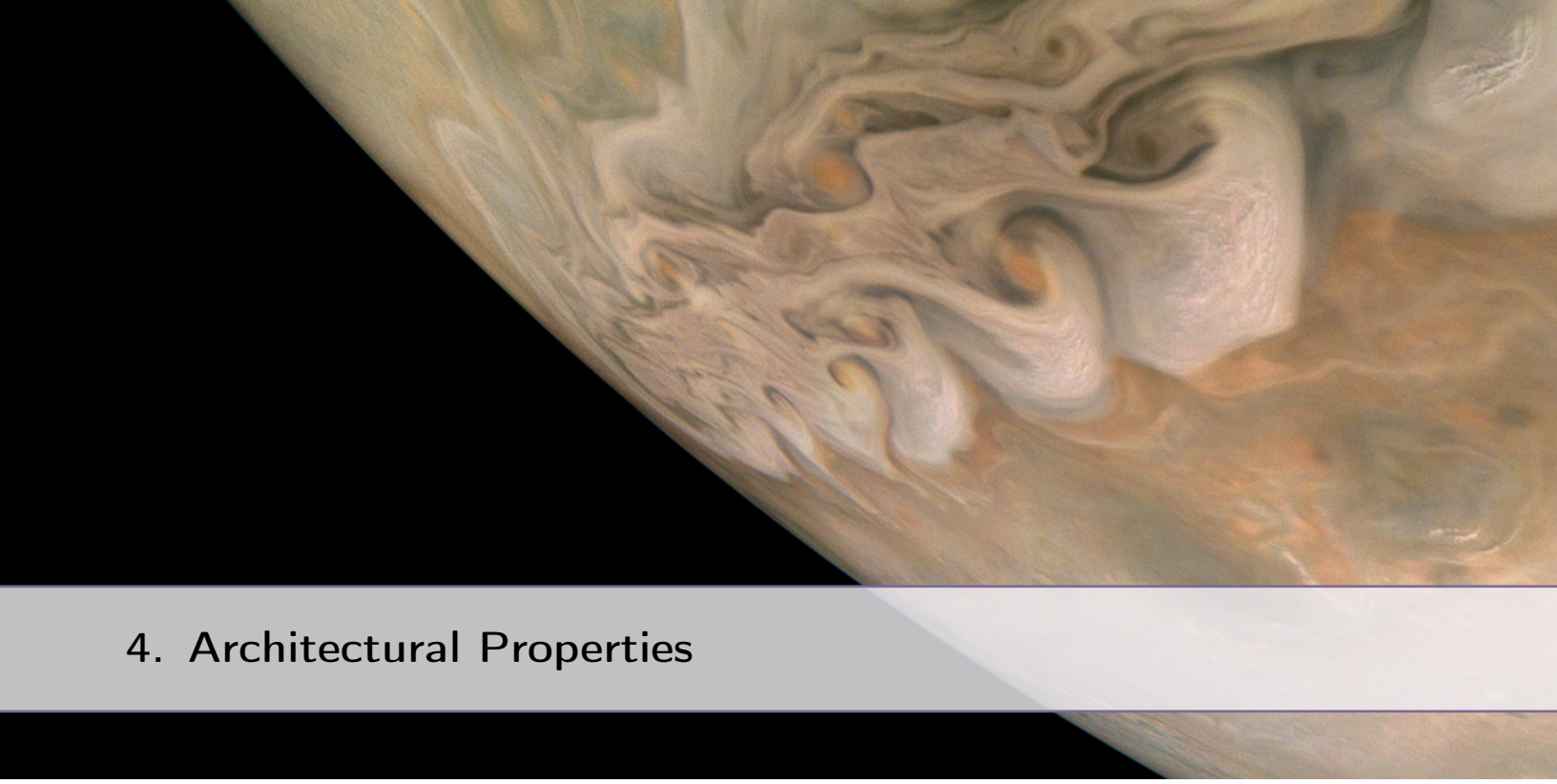
Outside of industry involvement, Nebulae concepts can be explored using existing NASA datasets and missions, first in a retrospective sense and then as an ancillary ground system pipeline. In particular, opportunities should be explored for missions that are in their extended life, since Nebulae-focused experiments would not interrupt core mission science. Efforts are already underway to explore implementing onboard decision making for missions in extended life, with NASA's Earth Observing 1 (EO-1) being a prime example in recent years. EO-1 was launched in 2000 as a one-year technology demonstration, with the mission successfully completed one-year later.[16] From the completion of its core mission through its eventual end of mission in 2017, EO-1 served as a pathfinder for implementing onboard algorithms for Earth Science remote sensing. Applications included hazard detection such as floods[17] or volcanic eruptions,[18] as well as cloud screening to prioritize data for downlink.[19] All of the onboard machine learning algorithms implemented on EO-1 had to comply with the severely constrained computing available on the spacecraft, which consisted of a Mongoose M5 processor running at 12 MHz and 128 Mb of RAM. Future missions should consider if small augmentations to required onboard computing or storage could enable extended-life Nebulae-like experiments, and subsequently greatly enhance the science return of the mission. With greater computing and storage resources available on board, a much larger variety of machine learning algorithms or data science approaches could be explored in later experiments.

[16] USGS. Earth Observing 1 (EO-1). `https://www.usgs.gov/centers/eros/science/earth-observing-1-eo-1`

[17] Chien et al. Monitoring Flooding in Thailand using Earth Observing One in a Sensorweb. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(2): 291-297. 2013.

[18] Davies et al. Observing Iceland's Eyjafjallajokull 2010 Eruptions with the Autonomous NASA Volcano Sensor Web. *Journal of Geophysical Research – Solid Earth*, 118(5): 1936–1956. 2013.

[19] Wagstaff et al. Cloud Filtering and Novelty Detection using Onboard Machine Learning for the EO-1 Spacecraft. *Proc. AI in the Oceans and Space Workshop, International Joint Conference on Artificial Intelligence*, Melbourne, Australia, August 2017.



# 4. Architectural Properties

In order to achieve Nebula's near-term and long-term goals, the system must be reliable, scalable, and durable beyond the lifespans of the missions it serves. To achieve this objective, Nebulae must be built with modular hardware components using standard protocols, and must expose well documented application-programmer interfaces (APIs) to allow instruments and scientific applications to consistently interact with the system. Abstracting the underlying Nebulae hardware will allow the capability to scale freely, and enable seamless upgrades and expansions as technology advances. Nebulae architecture, therefore, is primarily defined by the types of systems it supports and how the systems interact, rather than the technical specifications of each system.

## 4.1 Modularity

Nebulae comprise compute, storage, and networking modules designed to support a range of science use cases. Modules can contain more or fewer resources, and can be commercial-off-the-shelf (COTS) or Radiation-hardened by design (RHBD), depending on what a given mission demands and the destination where it will be deployed. For example, the Kepler space telescope might leverage a 100-TB block storage module, while MRO might only require a 10-TB module. In this example, regardless of the size of the block storage module, or whether it is a single drive or a RAID, the application will store and retrieve data using the same API. Modules can be plugged together using a standard protocol such as USB in order to achieve instant compatibility and reduce costs.

## 4.2 Scalability

Early on, Nebula-enabled missions will be self-contained, as they are today. However, one of Nebulae's goals is to repurpose and leverage resources on spacecraft long after their primary missions are complete. To achieve this goal, inter-spacecraft communications must be available even on early missions. To this end, all Nebulae-enabled spacecraft will be equipped with a high-speed optical communication system capable of sharing resources with multiple adjacent spacecraft. In the early phases, this will allow nearby spacecraft to share data with each other in order to augment their own datasets, in addition to the obvious benefits of bent-pipe relay to Earth. Additionally, spacecraft become nodes in a network that can relay information between a spacecraft and Earth when they are not in direct line of sight. However, the long-term benefits of creating a mesh network among spacecraft are much greater: as spacecraft approach end-of-life for their primary missions, they can be repurposed as external storage and compute servers. This would enable a nearby spacecraft to gain compute and storage resources on demand to execute expensive calculations without the need to transfer all of its data back to Earth. As the number of Nebulae-enabled spacecraft increases, ground stations—including the Deep Space Network—would need to expand and upgrade their capabilities, such as supporting optical communications.

## 4.3 Reliability

It is vital that Nebulae resources remain consistent in their operational capabilities. Programs must execute predictably and deterministically, and any discrepancy must be remediated or reported back to the application. Therefore, fault tolerance must be built into every aspect of Nebulae, from the individual modules (whether in hardware or software) to the APIs. Importantly, while fault detection is a must, remediation is not. It may be sufficient in many cases for a fault to be reported, but not necessarily addressed automatically by the system that faulted. For example, a process that detects a single event upset (SEU) while reducing data may simply halt and report an error back to the application. The application can then decide whether to retry the process or ignore it and move on. More serious faults such as single event latch-ups (SELs) or SEUs that fail without reporting an error need to be recognized and addressed by the overall Nebulae reliability architecture. Some SELs and SEUs need to be remediated quickly in order to prevent permanent damage up to loss of the mission. While it is vital that Nebulae hardware components are reliable, it is also important that software applications are reliable. Because Nebulae will be built around standard APIs, application software can use emulated modules hosted in the cloud on Earth during development with the expectation that the functionality will work the same in space. The increased communications capacity of Nebulae will also allow software to be patched and upgraded throughout the mission lifetime. This approach will enable software to become increasingly a primary mission asset.

## 4.4 Durability

Nebulae components must be reliable at least as long as their primary missions require, but in order to fully realize the concept, Nebulae must be designed to last as long as possible. After the primary mission expires, Nebulae may continue working in the same capacity, or they can be repurposed to share resources with nearby spacecraft to augment other missions. Durability may include everything from ensuring the hardware does not degrade (or degrades predictably over time) in the harsh environment of space, to ensuring data is not lost over time. Radiation tolerance is a necessary component of such survivability and must be matched effectively to the operating environment. A key aspect to ensuring data durability will be data cyclers that periodically back up data collected by spacecraft and offload it to Earth. In the event that a spacecraft loses a significant portion of its data (either via a software failure or a recoverable hardware fault), a data cycler can restore from a backup.

## 4.5 Upgradeability

As previously noted, it is expected that technology will continue to advance after the early Nebulae missions are launched, but it is crucial to the success of future missions that the early platforms do not become obsolete. In order to prevent obsolescence, and to embrace and utilize cutting-edge technology, newer spacecraft may use upgraded underlying technologies so long as they continue to support legacy interfaces. For example, in the event that 512-core rad-hard GPUs become available in a Nebulae compute module, while it will have significantly more data reduction power than a RAD750, the science applications will not have to change because they will access all compute modules utilizing a generalized API. If the API is allowed to change over time, the system necessitates a first-class version management system that embraces legacy APIs—particularly those exposed as shared resources—or requires certain missions to act as adaptors between legacy and contemporary versions. The latter is not recommended because it creates bottlenecks in the system when newer spacecraft need to communicate with older ones. Importantly, Nebulae's primary mode of upgrading will not be to upgrade existing spacecraft hardware, but to add new, more advanced spacecraft to the network, and to continuously upgrade software. In the future, it may be possible to perform on-orbit (or in-flight) module swaps (perhaps a function of a data cycler or specialized maintenance spacecraft) or even to 3D print modules with advanced capabilities in situ. While a fundamental concept of Nebulae is to continue to utilize spacecraft as long as possible, it will be critical to avoid imbalance within the system. Eventually, as hardware degrades or the technology is no longer capable of keeping up with newer missions, older spacecraft may be significantly overtaken by newer spacecraft. The best antidote to imbalance is modularity, but even if interfaces continue to work as intended, the contributions of older Nebulae elements (e.g., for computational throughput) may become so stressed that they should be retired, or transitioned to interested academic institutions.

## 4.6 Organization and culture

Perhaps the most difficult aspect of Nebulae's architecture to achieve will be shifting the culture of mission design. Starting with the earliest missions, it will be vital for mission designers to think about the possibilities offered by additional compute and storage capabilities (even short of machine learning) instead of focusing on the minimally required system for achieving specific science mission objectives, which are inevitably overtaken by discoveries. Organizational change is easiest when it is embraced and modeled at the top. If senior NASA officials embrace a culture of incremental improvement and multi-mission cooperation, mission and system architects will be more willing to embrace such a shift themselves. It is recommended that the agency consider incentivizing mission designs that implement Nebulae in some capacity in order to stimulate its adoption. It must also be emphasized that Nebulae are intended to reduce costs over time, leading to more science per dollar. Nebulae architecture works to reduce cost and boost scientific discovery by empowering missions to **focus solely and boldly on the science**, and not the increasingly sophisticated (but largely invisible) plumbing that makes it all possible.

# 5. Pragmatics—The Engineering of a Nebulae System

All is naught if the design cannot be realized. The Pragmatics subteam of the KISS Workshop on Nebulae was comprised of nearly half of the participants, and investigated current relevant technologies and architectures, extrapolated technology growth, and provided examples of components and subsystems necessary or useful to create Nebulae-enabled spacecraft. What would a Nebulae system supporting a variety of missions look like?

The challenge of every deep space mission is to meet data science processing requirements while minimizing Size, Weight, and Power (SWaP) to the greatest extent possible. The harsh radiation environment is also a key driver that must be addressed.

As noted earlier in this report, processing data remotely is considered secondary to, if not anathema to, current science paradigms because the full results cannot be reproduced on Earth. However, the task of Pragmatics subteam is to enable this additional flexibility and let the scientific community utilize it to whatever degree they wish on each mission, or during different phases of a mission.

We focused on identifying modular Nebulae components and building blocks that can be configured to support Mars missions, Jupiter and Saturn moon missions, and future astrophysics missions, as described in the Science Motivations and Opportunities section.

We examined key engineering technologies that will drive Nebulae's overall performance, and looked at the market trends in each area compared to the anticipated needs of a Nebulae system.

Focus areas were:

- Onboard processing capability
- Onboard data storage capacity
- Software and data architectures
- Resiliency and fault tolerance architectures
- Spacecraft communications and networking
- Power generation

Nebulae computing assemblies are proposed for three exemplar missions reflected in Figure 5.1. We conclude with recommendations for technology maturation investments to address technology gaps. A technology gap is any technology need which the commercial market does not appear to be resolving in a timely manner, and may indicate a need for additional stimulus from the space community.

What is different, from a spacecraft perspective, between the computing system design for a conventional deep space mission and for a Nebulae-equipped mission?

- **The basics are the same: Minimize the power requirement, weight, and size.** Size and weight drive launch costs and also impact fuel requirements. Fuel constraints set the limit for maneuvering and station keeping which can limit the mission's operational lifetime. And generating power in deep space is extremely challenging and expensive and is often the most constraining factor in the final design.
- **The challenging physical environment for compute and storage is the same.** Radiation impacts the selection of components and the detailed architecture and design of circuits. Extreme environmentals for cold/heat thermal cycling impact circuits and cards, and shock and vibration from launch and maneuvering are the same as always. However, when Nebulae nodes utilize circuits that are not fully radiation hardened, new packaging and shielding will be employed, and new software will be utilized that allows for recovery and reset of radiation disturbances.
- **The core vehicle health and control system still must be highly reliable.** The processor(s) and storage comprising the root of the control system cannot fail. They must be fully radiation hardened. Secondary processors can be rebooted or reset as long as the mission data is protected appropriately.
- **Key difference: Nebulae need dramatically more processing and storage capacity because the concept moves some of the Science Data Processing and**

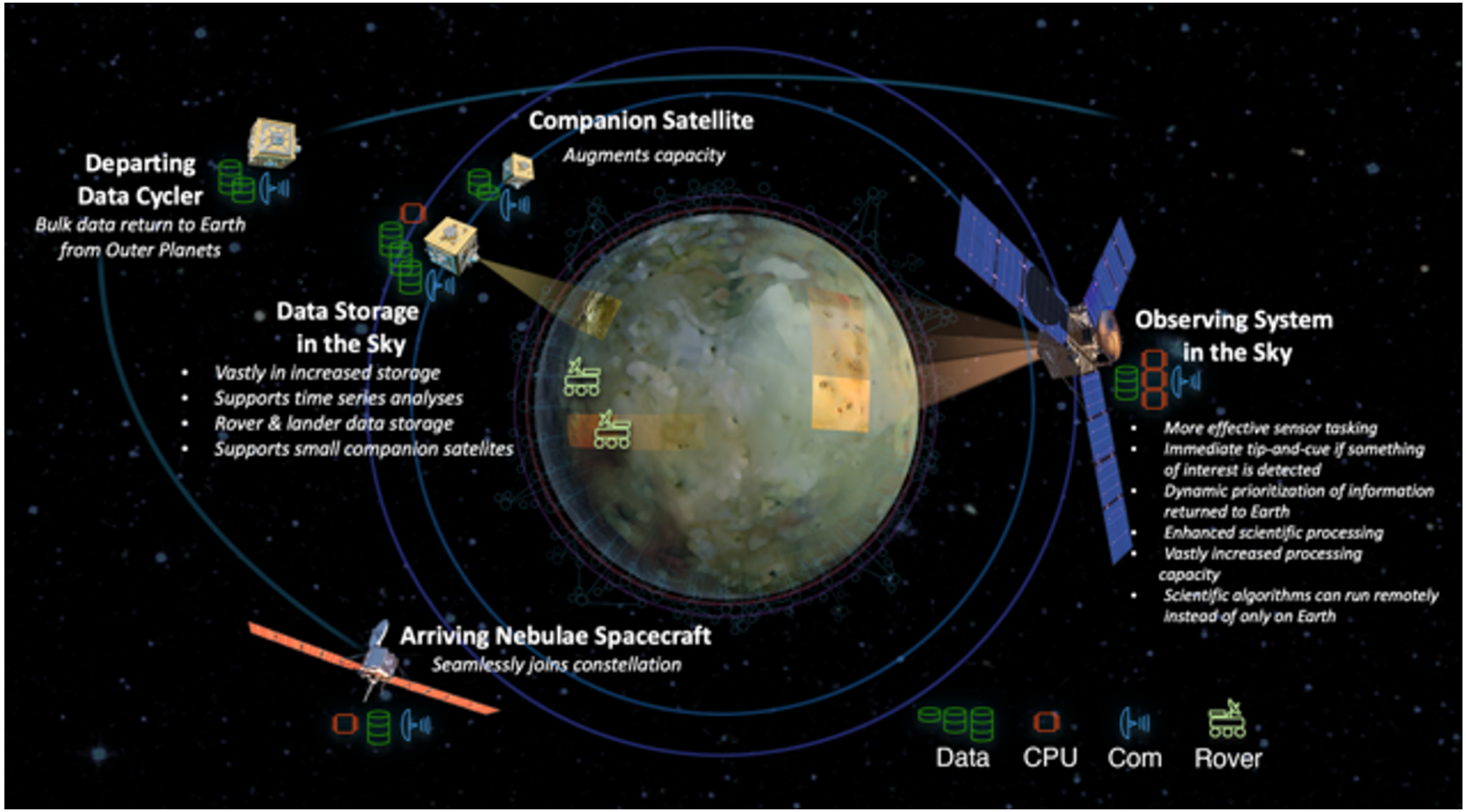



Figure 5.1: Nebulae can enable science missions to perform additional in-situ processing by providing additional modular storage, processing and/or communications.

**Storage (SDPS) from Earth-bound to on board the space platform.** In order to execute mission processing on board, filter data, etc., requirements for onboard computing and storage must increase to a scale not seen before. Processing demands increase 1 to 2 orders of magnitude. As described earlier in this report, storage demands are insatiable at the extremes, as sensors generate PBs of data on a recurring basis. While not all data can be stored indefinitely even in a Nebulae configuration, storage demands will increase at least 5× and up to 100× to enable caching of enough data to run in-situ analyses over greater surface areas or longer time series. Nebulae-enabled missions support a new scientific paradigm allowing for remote processing and potentially some raw or intermediate data not being returned (immediately) to Earth.

- **Key difference: Nebulae likely uses a heterogeneous multiple-processor architecture.** The team advocates the use of a mixture of fully radiation-protected chips and non-fully radiation protected chips and boards, dramatically reducing cost per onboard operation and allowing the use of previously Earth-bound higher performance processors. The architecture also enables the use of special purpose mission processors such as a neural network accelerator, a processor optimized for matrix mathematics and machine learning algorithms, or a high-speed LIDAR processor supporting autonomous flight or more complex autonomous landing.

Thus, from an engineering perspective, how do we meet these new requirements while still addressing the long-standing ones? We tackled the problem by examining capacities and trends and modern architectures, and then tying them together in a scalable, resilient canonical solution.

## 5.1 Market Trends vs. Requirements—Spacecraft Processing Capacity

The primary characteristics of a Nebulae node are determined by its **processing capacity and its fault tolerance architecture**. With sufficient processing capacity, analysis could take place in situ at a rate such that the sensors could be adjusted in near-real time, optimizing the collection of meaningful data. With sufficient processing capacity, multiple sets of multi-stage scientific processing could be performed in-situ, converting data into information that requires much less bandwidth to transmit back to Earth.

Capacities are of paramount concern to the **Data Server in the Sky** and **Data Cycler** Nebulae node classes.

### 5.1.1 Processor Capacity—CPUs

RHBD CPUs are approximately 2 to 3 technology generations and 100× behind commercial state-of-the-art CPUs in processing capacity. We suspect this will always be the case. (See Appendix A.)

A key recent technological advance which enables Nebulae is the emergence of multi-core CPUs. The gap between commercial and RHBD processors has widened since the release of the BAE RAD750 processor in 2003. RHBD single-core processor capacity has stagnated. However, the emergence of multi-core RHBD processors reduces the gap to commercial processors more towards historical norms.

The most important impact of multi-core CPUs is that processing capacity is moving on board in the commercial and government markets. As mentioned before, the root and core of the vehicle's control system—the command and data handling (C&DH) subsystem—cannot fail. C&DH software is rigorously tested, rarely altered after launch, is partially or fully written in hard-real-time methods, and is isolated from sensor processing and other software. C&DH control software was allocated one entire single core processor. The advent of multi-core RHBD processors (4, 8, and soon 16 cores per processor) means that the C&DH can be protected and isolated to a core, and now there are 3, 7, or even 15 additional cores of capacity available for the mission, essentially free from a SWaP and cost perspective. This motivated the commercial satellite market to create layered control and application architectures that can take advantage of the new capacity. These new architectures directly enable Nebulae's multi-processor capabilities and in-flight mission processing upgrade capabilities.

Meanwhile, processors for business and analytic functions have many more cores, faster clock cycles, and higher-density circuits that could perform spacecraft science much faster if they were not incompatible with the space environment.

Therefore, **while the Nebulae core C&DH processor and spacecraft "watchdog" functions will always reside on a RHBD processor, a fully RHBD Nebulae node will always be more expensive (and perhaps unaffordable) compared to a Nebulae node architected to take advantage of a mix of processors.** We looked at ways to increase compute power by trading inherent radiation performance for processing capacity, and addressing radiation impacts in other manners.

### 5.1.2 Processor Capacity—GPUs, DSPs, FPGAs

We considered other recent advances in processor design such as Graphics Processing Units (GPUs), Digital Signal Processor circuit cards (DSPs), and Field Programmable Gate Arrays (FPGAs). Each of these processor types can provide a dramatic improvement in speed and capability per size/weight/power allocation, if the algorithm is closely matched to that processor. Unfortunately, radiation-hardened versions are not usually available, and if they are, they lag several generations behind the commercial or rad-tolerant versions.

FPGAs such as the Virtex and Ultrascale processors have already been flown in spacecraft in rad-tolerant payload processing configurations. DSP cards and chips are also an alternative which may be best aligned with a particular algorithm. Both of these are relatively difficult to

program and are not meant for frequent reprogramming during operations. FPGAs and DSP cards have a long history of spaceflight operations and will continue to be used for special purposes on Nebulae nodes, such as the collection and initial processing of data from high throughput sensors.

The Graphics Processing Unit (GPU) and recent GPU evolution offer new and enticing possibilities because of support of machine learning (ML) and deep learning (DL) artificial intelligence software including computer vision applications. GPUs are available with significant programming toolkits and are designed to support a wide variety of software evolving over time.

Graphics processors originated to efficiently process real-time visualizations required for video games and for high-performance renderings of complex images. GPUs are dramatically more efficient than general purpose CPUs for certain classes of algorithms, including graphics processing for machine vision, signal processing, machine learning, and matrix manipulation. However, no GPUs are RHBD today or in the foreseeable future. However, there is great possibility in leveraging GPUs for additional compute power given their easy reprogrammability and unmatched efficiency in algorithms of interest (e.g., deep neural networks) to remote uncrewed spacecraft.

**Layered software architectures** evolved to support multi-core processors and are extensible to support a heterogeneous mix of processors. Nebulae will use a layered software architecture to efficiently support a mix of CPU and special-purpose processors. A well-chosen architecture allows for mixing and matching of processors across space vehicles such that the processors can be maximized for each vehicle while isolating the rest of the system from change and maintaining interoperability across disparately implemented Nebulae nodes.

Machine learning algorithms, deep learning algorithms, neural networks of most types (including neuromorphic implementations) have all been shown to execute significantly more efficiently and rapidly on special purpose processors than on general purpose CPUs. The auto industry in fact is a primary driver in the production of small, rugged, inexpensive GPUs in order to support self-driving vehicles, and auto-qualified parts are increasingly making their way into space vehicles.

We assert that Nebulae nodes will be much more efficient if they can leverage GPUs for some in situ sensor processing and data reduction. Sensors may also have embedded DSPs or FPGAs that will provide optimized processing for a fixed set of algorithms and will not be part of the Nebulae dynamically reprogrammable deep-space computing resource pool. The proof will be provided by characterization and measurement of actual scientific algorithms, as described further in Appendix C. Nebulae will likely use GPUs selected for radiation tolerance which will be protected with additional radiation shielding.

Modern fault-tolerant architectures and layered software architectures allow for the restart and recovery of specialty processors and secondary processors if they are disrupted. Data can either be reprocessed if critical or skipped if not critical.

### 5.1.3 Processing Demand

Having science data for a mission remain onboard rather than being downlinked to Earth changes the model for how the data would be collected and how humans would access them. For the purposes of this subsection, we will use the following notional architecture of how a mission with an orbiter ("Mother") and lander ("Daughter") would be used to provide Virtual Science Data Center operations to users. The orbiter has an onboard archive, which is the repository for almost all science data collected.

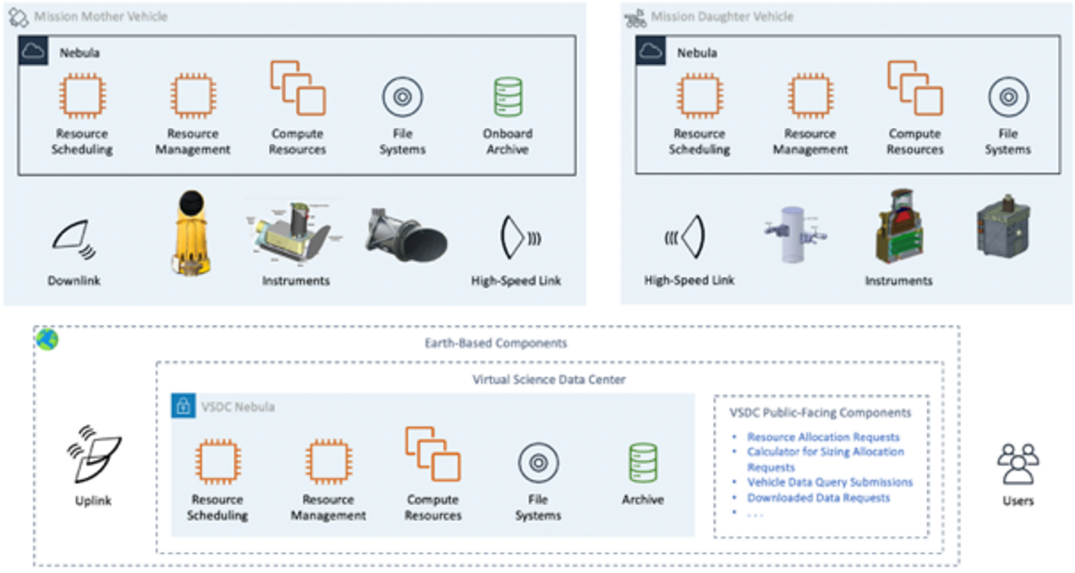



Figure 5.2: Models for data collection and processing demand.

While humans on the ground would have control of the mission at a strategic level, communication latencies will preclude human-in-the-loop decision making at a tactical level. Following prioritization instructions from human controllers, Nebulae will direct instruments for data collection, analyze the data, and prioritize the return of data products to Earth.

For example, Nebulae might be instructed to give high priority to the analysis of data from a low-resolution instrument to detect if changes have occurred from the baseline data. When the analysis software detects an interesting change, it could initiate a high priority activity to start collecting data from a high-resolution instrument while that surface region is still within range. The autonomy capabilities of the mission are critical in this example because involving a human in the decision would miss the opportunity. These measurements need not interfere with the usual "Transactional" uplink/downlink or "Remote Instrument" model of operations.

Besides enabling a degree of autonomy with respect to data collection, Nebulae also needs to serve as the primary science data processing engine. This could take the form of humans supplying Nebulae with a prioritized list of processing tasks. Some of them will be normal housekeeping tasks such as are normally found in a science data pipeline. Others might be "database queries" or deep analysis tasks that have been formulated by a ground team in order to answer a question posed by a researcher.

One of the roles of the ground team is to minimize the processing and communication time needed by mission resources to handle the query. Using previously downlinked data as much as possible would be an obvious expedient. Similarly, downlinking an intermediate dataset (or scheduling for a Data Mule transfer) may make sense if subsequent processing won't further reduce the size of the output. Taking a collection of queries and looking for common intermediate data products would also help.

## 5.2 Characterizing the Compute Workload

Nebulae applications can be characterized using methods outlined in Appendix C. Scientific applications range from 1–10 GOPS (billion operations per second) for autonomous mission planning to 10–50 GOPS for fast traverse and landing to over 50 GOPS for radar and hyperspectral imaging. From these characterizations we can estimate the compute resources required and select a mix of processors that achieve that capacity cost-effectively and at the required level of reliability, per the Reliability as a Service architecture explained in Appendix D.

The compute resources on a Nebulae-enabled mission do not stand alone. They compete for power with other systems such as communications. Computing may be throttled back during times of very high power demands; or may be able to utilize surplus power at other times. In a sense, power is a resource to be scheduled, just like communication time, instrument time, archive time, or compute time.

**For a medium- or large-sized Nebulae space platform that will be performing heavy sensor computing, we prefer one core fully RHBD processor for C&DH and spacecraft "watchdog" functions, plus others if a high-availability configuration is required (such as for human spaceflight), plus a set of at least two adjunct sensor processors, at least one of which would be a GPU for efficient machine learning and matrix calculations.**

### 5.2.1 Data and Information Storage Capacity

The amount of fully radiation hardened volatile RAM and non-volatile RAM (NVRAM) required for spacecraft flight control remains relatively constant, as do the compute requirements. Additional sensor processors require additional RAM. Caching and archival of sensor data and

data being processed requires a much larger amount of RAM and NVRAM—as much as can reasonably be provided.

Radiation-hardened memory (non-volatile storage) lags 10–20 years behind commercial technology, equivalent to about 2 generations of technology or 2–3 orders of magnitude of capacity. A comparison of commercial and RHBD memory capacity is provided in Appendix A.

Radiation-tolerant storage technology currently can provide about 1 TB of solid-state storage in a size about the same as an iPhone. Deep space missions in 10 years should be able to obtain 100 TB of solid-state storage suitable for a 15-year mission, in roughly the size of 3 shoe boxes, running on 50 W of power.

We extrapolate that 10-GB RHBD modules are available in 2030 for our Nebulae "what if" scenarios. **Thus our basic configuration would provide about 10-GB RHBD RAM for the core processors, plus 10–20 GB of rad-tolerant RAM for auxiliary processors. Additional volatile and non-volatile storage would be provided until the limit of available size, weight, or power was reached. Non-volatile storage—the main storage for the mission—would ideally reach 100 TB or more on medium-sized vehicles.**

### 5.2.2 Radiation Impacts, Reliability, and Resiliency

Radiation impacts electronics and circuitry, causing recoverable and irrecoverable failures. The space environment is defined by three sources of radiation: Galactic cosmic rays, solar radiation, and radiation belts.

Spacecraft components can fail due to the long-term cumulative dose or by single high-energy particles that impact a tiny part of a chip, causing a single-event upset (SEU).

There are four primary avenues for reducing or preventing radiation-induced failures:

- Using radiation-hardened-by-design (RHBD) processors specially produced to withstand deep-space environments. RHBD electronics are very reliable and SWaP-efficient.
- Shielding electronics from radiation with metal enclosures. This technique enables the use of non-RHBD electronics that are more powerful and capable than RHBD devices, at the expense of size and weight, and imperfect shielding.
- Testing and screening commercial components and selecting parts that exhibit radiation tolerance. These parts are less tolerant than RHBD electronics, but may be sufficient for a particular mission. These so-called "qualified COTS" are more prevalent in recent missions, such as experiments with the Snapdragon processor from Qualcomm.
- Architecting a system that can detect radiation-induced faults and recover from them, using multi-processor fault tolerant architectures and layered software infrastructure.

Deep space exploration requires computational resources beyond what can be provided solely by state-of-the-art RHBD processors. Either we need to produce RHBD processors that are one to two orders of magnitude more capable—with new releases every decade—or we need to utilize the other three avenues available to us.

Nebulae will offer **Resiliency as a Service (RaaS)**—a means to leverage all available avenues of fault protection, fault tolerance, and radiation protection to provide the reliability and availability needed for scientific missions in deep space.

### 5.2.3 Resiliency as a Service

The resiliency of a system is measured using a combination of hardware characteristics derived from wiring diagrams and component failure rates together with software resiliency characteristics such as implementations of redundancy, automated failover, and recovery. Higher resiliency nearly always costs more than lower resiliency in some manner: hot spares are essentially wasted capacity; RAID solutions consume more storage space per bit of data than non-RAID; watchdog processors and software consume power and space without providing direct mission value.

Traditionally, computational capacity and redundancy are static calculations. If the software characteristics are altered—let's say the data in storage was implemented in a RAID5 configuration—then the resiliency of the system is altered. We recommend a new concept of Resiliency as a Service whereby the processing subsystem and storage subsystem's resiliency attributes are scalable, adaptable, and most importantly dynamically adjustable in order to make science data processing and storage fault tolerant.

To dynamically change the Resiliency posture:

- The N×M redundancy scheme of multiple CPUs can be changed on the fly.
- Hypervisors provide restarting of non-mission-critical processors.
- The RAID posture can be altered, and data stores can be configured independently.
- Layered software abstracts the CPUs from the applications, allowing for movement of programs to different resources for efficiency and resiliency.
- Software provides checkpoint/restart functions for failure/fault recovery.
- Software provides the ability to re-process scientific data if a failure occurred during processing.
- Networked cloud capabilities support processor-agnostic software, uploadable new algorithms, and software-based resilient behaviors.

RaaS requires multiple processors and cross-connected storage and I/O, plus a well layered software architecture, plus some additional emerging software technology.

Proposals for RaaS fault tolerance, layered software architectures, and handling of radiation-induced faults are discussed in Appendix D. What is important for Nebulae is to keep as many options for dynamically adjustable redundancy available when selecting hardware and software.

### 5.2.4 Data Standards

To facilitate interoperability between Nebulae nodes, it is imperative that enabling data standards be adopted early in the architectural development process. Making the right choices will result in an enduring capability that not only meets the near-term mission needs but embodies qualities of flexibility and extensibility that will allow the flexibility for system capabilities to evolve over time, much like the Internet Protocol standard has allowed the build-out of an incredibly rich and diverse ecosystem of protocols hosting world-changing technologies that could never have been imagined in 1974 when the original standard was defined.

Data standards support the re-processing of raw data and provide for tracking the progeny of processed data. As more data is transformed in situ, each alteration of a dataset must be precisely defined to avoid unintentional misapplication or misuse by a software application.

The data standards in need of definition for the Nebulae architecture include:

- An "on the wire" messaging standard describing the means by which data is encoded in a single datagram transaction. This will describe how the data is formatted, such that both the sender and the receiver of the message understand how to interpret its contents. Ideally, the description of how messages are formatted should be defined by a metalanguage that provides a high degree of flexibility in how data is represented. That standard would not just be constrained to identify the type of data item (a primitive such as a floating point of a certain width) and its location in the message. The standard would also allow for complex organization of primitive data items into structures and hierarchies—to provide a significant degree of flexibility to accommodate the transport of highly dynamic data that is dependent on conditions at the time of message generation. One can think of this approach as the class paradigm of object-oriented programming, which builds on dynamic data constructs such as strings to define layer-upon-layer of inheritance to arrive at a top-level representation that potentially derives from many sub-definitions.

- A schema, which in addition to defining the format aspects of each message transported, also defines the sequencing of messaging that must be performed in order to achieve a certain result between two architecture elements. In some cases, messages can simply

be delivered in publish-subscribe fashion, where one component asserts its desire to receive information of a certain type from another component and subsequently receives that message periodically or on an event-basis. Other cases call for the use of a service model, where a message is sent to another component and a reply is expected back. Other patterns also exist, such as one-way messaging initiated by one component to another (commands). The Nebulae architecture should support all of these interaction models—the combination of all facilitates a richness of interoperability that will meet the broad range of mission use cases expected. Middleware solutions such as the Robot Operating System support these various interaction patterns and should be looked to for inspiration on this front.

- An ontology, defining the "meaning" of data included in messages. An ontology can also be thought of as a dictionary of terms. It is not sufficient to simply understand that a data item is of a certain type. More importantly, that data item must be relatable to a particular context (e.g., what it represents). Typically, this information is held in interface control documents and is interpreted by humans when they write custom code to extract data from messages and then connect them to logic functions. In a Nebulae implementation it is desired to have an electronic ontology perform this function. Different initiatives have taken a variety of approaches toward structuring ontologies for their systems. Most employ a hierarchical approach of some kind where name spacing is employed to organize "the dictionary" in a logical way (e.g., $nebulae.asset.satellite.subsystem.attitude.bodyRate.x$). Whatever the means by which the ontology is represented, it should allow for unambiguous description of data, so that any developer or user of the system can properly interpret its meaning and intent. With a well structured ontology comes the ability to employ a variety of approaches that will be instrumental to revolutionary and/or evolutionary Nebulae features such as rich data searches, data mining, and automated code generation in support of processing algorithms.

Prudent selection of technologies that embrace the above three categories should be informed by current state-of-the-art in terrestrial cloud computing. However, many of the existing solutions have not been developed with constrained computing environments in mind—they target large enterprise systems with plentiful processing and memory resources. The right solution will be one that possesses the desired characteristics in balance with a very frugal and deterministic approach to the use of computing resources, which will be absolutely essential to reliable and very-long-duration performance on nodes that will be extremely distant from Earth and thus not routinely upgraded.

The difficulty in changing data standards after deployment necessitates that these choices be made very carefully—they are essential to the ultimate success of the Nebulae concept.

The data standards must be evaluated and selected with a lens of well-posed use cases and against well-stated capability criteria.

## 5.3 Spacecraft Communications and Networking

Communication to and from spacecraft in deep space today is accomplished with NASA's Deep Space Network (DSN) and similar capabilities from other nations' space agencies. The DSN provides a set of large communications stations located at three sites around the Earth so that at least one of these sites maintains line-of-sight connectivity with a spacecraft at all times (unless the spacecraft is occulted from Earth by another body or disrupted by terrestrial weather.)

Though this system can provide continuous communications with a spacecraft in deep space, it is typically not used in this way. Instead, communication sessions are scheduled in advance with each spacecraft. During these sessions (which might be daily, weekly, or sporadic, depending on the data needs of the particular mission) commands or programs are sent from Earth to the spacecraft while engineering and science data are returned to the Earth.

DSN (and like capabilities) are continually upgraded to keep pace as much as possible with the demands of the science missions in deep space. Improvements to the DSN are in process or planned that will increase the basic capability (measured in bits/second returned from deep space) by approximately a factor of ten over the next decade.[1] This rate of improvement is expected to continue.

We do not propose to change this paradigm. Rather, we will use it as the preferred method of communications between the spacecraft and the Earth. However, the Nebulae concept will allow "information return" from deep space to grow much faster than the "data return" projections of an order of magnitude each decade. The reasons for this claim have been explained above.

Today, missions schedule the set of data they expect to return during each communications session. This will change as the DSN and spacecraft implement more advanced networking technologies such as Disruption/Delay Tolerant Networking (DTN.) DTN is an evolving international networking standard that extends terrestrial Internet service concepts into deep space, while taking into account the differences between terrestrial and deep-space communications (most notably, the light travel time). Data will be transmitted between a spacecraft, possible communications relay elements, and Earth in such a way that there is assured end-to-end performance and completion. Any bits that fail to reach their destination correctly during a specific communications session will be automatically retransmitted in subsequent sessions. This will lead to autonomously-optimized sessions.

[1] Deutsch, L. J., Stephen A. Townes, S. A., Liebrecht, P. E., Vrotsos, P. A., & Cornwell, D. M. 2016, "Deep Space Network: The Next 50 Years," *SpaceOps 2016 Conference*; doi: 10.2514/6.2016-2373

DTN is being infused into the evolving communications architecture for international lunar exploration, sometimes referred to as "LunaNet." LunaNet derives from recommendations published by the Interagency Operations Advisory Group (IOAG) and is being supported by most of the world's civilian space agencies. As LunaNet is deployed for missions including Artemis, we will glean experience that will be valuable to developing Nebulae.

### 5.3.1 Data Cyclers

An intermediate option for enhancing the science return of future missions that was explored during the study was that of "data cyclers." Rather than all of the data being processed "in the (deep-space) cloud" or being transmitted to Earth via telecommunications links, some or all of the data could be moved physically from the remote target object to close to the Earth. This section summarizes the results of a trade study conducted during the course of this workshop comparing the costs and benefits of these various approaches to transferring data from a target body to the Earth.

**Data Cycler:** One or more spacecraft that are placed on transfer orbits between Earth and the target body. When a data cycler is close to the target body, it receives data from the asset or assets at the target body, where presumably relatively high transfer rates can be sustained due to the short ranges. The data cycler simply stores the data until it approaches the Earth, where again the short transfer ranges allow high data transfer rates. The architecture is analogous to techniques used on the Earth for which relatively high data rates can be sustained simply by transporting physical storage media (e.g., hard drives) from one location to another (often termed "sneakernet"). The spacecraft is optimized for data storage, with only minimal other functionality. In particular, if the spacecraft is on a (Hohmann) transfer orbit between Earth and the target body, only minimal propulsion would be required. Optionally, the cycler can continuously transmit (or receive) data while in transit if it contained a longer-range comms link—this combination of relay and store-and-forward might increase by terabytes the amount of information returned by the cycler each trip.

**Data Relay:** One or more spacecraft are placed intermediate to Earth and the target body to relay data. Because the distance between the target body and the relay (or relays) is smaller than that from the target body to Earth, higher data rates can be sustained. In this case, the spacecraft would be optimized to have the most capable telecommunications system, with a telecommunications system comparable to or exceeding the capabilities of the Cassini spacecraft as a useful benchmark (~100 W transmitter, 4-m antenna capable at both X- and Ka bands). Breidenthal

(2000)[a] has considered this scenario in depth, including multi-hop relays with the potential for continuous coverage to a target body. As low-cost optical terminals emerge over the next decade, the additional capacity per Watt may significantly improve the business case associated with relays. Relays at the Moon and Mars and L1 point have a different business proposition because we anticipate multiple landers, probes, and other orbiters in continuous operations on/around each body as the 21st century progresses.

[a]Breidenthal, J. C. 2000, "The Merits of Multi-Hop Communication in Deep Space," in *IEEE Aerospace Conference*; doi: 10.1109/AERO.2000.879389

### DSN 34m antenna

As a data transfer metric, we consider a DSN 34-m antenna (these are a standard feature of NASA's Deep Space Network and are used to transmit commands and receive telemetry from spacecraft from across the Solar System) that receives science data (telemetry) at a rate of 1 Mbps during weekly tracking passes that have a duration of six hours. For reference, telemetry from the Mars Reconnaissance Orbiter (MRO) can be obtained at the rate of 6 Mbps, when Mars is near opposition, using its X-band system. MRO also carries a Ka-band system that was intended to demonstrate telecommunications at that frequency band, and it could, in principle, obtain data transfer rates as high as 25 Mbps. Over the course of six months, a series of DSN 34-m antennas (hand-off as the Earth rotates) could receive a data volume of about 560 Gb = 70 GB (= 1 Mbps $\times$ 6 hr/week $\times$ 26 weeks) from one space vehicle. A data cycler or data relay has to enable a comparable data volume to be obtained over a similar interval, with six months being the approximate duration of the one-way orbital transfer from Mars to Earth; data cyclers to the outer Solar System would have longer orbital transfer durations and therefore would be required to transport even larger data volumes.

For purposes of the trade study, we considered a financial benchmark of $60M, as this figure is close to both the cost cap of a Small Innovative Missions for Planetary Exploration (SIMPLEx) mission (excluding launch cost) and close to, if somewhat below, the construction cost of a DSN 34-m antenna. For a SIMPLEx-class mission, an ESPA-class spacecraft, with no more than180 kg in mass (wet), serves as a useful reference.

The essential challenge of either a data cycler or a data relay is now clear. For approximately the same amount of funding as for 2 or 3 cyclers, it is possible to obtain a series of 3 DSN 34-m antennas that could support and enable multiple missions across the Solar System. Further, with reasonable maintenance, a DSN 34-m antenna could operate for at least 20 years, and potentially up to 50 years. By contrast, a data cycler can enable missions at only one target body and usually for no more than 2 passes. Further, the longevity of ESPA-class spacecraft in deep space is unclear.

In order to be a worthwhile cost-benefit trade, a data cycler or data relay architecture would likely have to make use of spacecraft that are at least an order of magnitude less expensive than current spacecraft and with a demonstrated lifetime in deep space in excess of 10 years. Extrapolations of optical terminal capabilities in range and bandwidth may further constrain the applicable envelope for cyclers. Look to the NASA Innovative Advanced Concepts program to continue this study and provide final conclusions.

## 5.4 Deep Space Power Generation

Power generation is critical for spacecraft. Power limitations often cap the communications to Earth, which in turn limits total mission science delivered. Communications data rates are directly proportional to the electrical power applied. Sensors and data processing units are the other primary consumers of power.

The power available to a radio or optical transmitter limits the maximum data rate and limits the amount of computing resources that can be utilized. In Earth orbit, there is usually sufficient power available via large solar arrays such that other factors impact and limit communications data rates. Beyond Mars, the available power is the primary limiter to spacecraft communications data rates.

Spacecraft electrical power is generated in one of three ways: via solar arrays, radioactive thermoelectric generators (RTGs), or small fission reactors. The United States has launched only one reactor, in 1961, and is unlikely to do so to power a spacecraft.

Radioactive Thermoelectric Generators (RTGs) are a mature technology in use since 1961. No significant increase in capacity or capability is foreseen in the near future. RTGs are predictable and reliable, but they are also inefficient, expensive, require nuclear material from the Department of Energy, and have a substantial regulatory burden.

Accounting for efficiency improvements in solar array technology, and including estimates from RTG improvements expected by 2030, in Jupiter orbit, RTG power provides (and will continue to provide) an equivalent overall power to a 60 $m^2$ solar array. Beyond Jupiter, RTGs provide significantly more power than solar arrays. Nebulae systems will be power limited as the distance from the Sun increases, and power efficiency will be a key constraint in the architecture, design, and component selection for Nebulae equipment.

As a point of departure and general rule, solar arrays are the most effective way to generate power in space for spacecraft in missions from the inner solar system and out as far as Jupiter. Rovers and UAVs require trade studies to determine whether solar arrays are effective or RTGs are required; Martian rovers have utilized both. RTGs are often the recommendation for deep space missions and large landers/rovers. Missions beyond Jupiter nominally will utilize

RTGs, except Data Cyclers which seem unlikely to be allowed to consume scarce RTGs just for several passes.

The initial calculations for a Nebulae power system will start with the available size and weight for the power generation system, determine which technology provides the most power given the mission profile and environmental constraints, and derive the maximum initially available power.

## 5.5 Designing Nebulae-Enabled Systems

Using the technical information and analyses provided in the Appendixes, we provide **3 sample Nebulae configurations** illustrating the driving cases discussed previously and illustrated in Figure 5.1. The configurations are driven by scientific need and bounded by the realities of the state of the art.

A. "Mars Reconnaissance Orbiter Prime": MRO Prime is an imagined next-generation MRO performing a similar **Data System in the Sky** mission. MRO Prime assumes a Mars-orbiting spacecraft with the same size and weight as MRO which performs its own science and also serves as a relay platform and compute server for rovers, landers, and other sensor spacecraft. MRO Prime has a science package, a Nebulae node for processing, and a heavy relay capability. The Nebulae node utilizes advancements in processor and storage technology, a blend of RHBD and qualified COTS hardware, and a layered architecture to provide on-site reconfigurable services to the mission. The new services and expanded processing and communications capabilities allow MRO Prime to produce additional data supporting enhanced scientific results. Generally speaking, the MRO Prime design is constrained by the size and weight allocation associated with the chosen launch vehicle.
    - In a footprint of roughly 0.5 $m^3$, a Nebulae node design in 2025 for a Mars-orbiting spacecraft might provide an RHBD HPSC chipset controlling the vital bus functions plus 5 CPUs and 2 GPUs for scientific processing and data reduction. Using Resiliency as a Service layered software, CPUs can be organized in a variety of redundancy configurations during critical functions such as near-real-time support of a lander or rover, or could be configured in a non-redundant arrangement for maximum sensor processing during routine operations. Depending on the processor selections, there could be at least 3 TFLOPS of scientific processing capacity.

    - The subsystem could have a non-volatile storage of 50–100 TB (persistent through power loss and restarts) and a volatile high-speed RAM capacity of approximately 40 GB.

- The subsystem would require approximately 1.5 kW of continuous power, part of the 8 kW generated by an 60 $m^2$ solar array.
- High-bandwidth communications with Earth would be provided via a 20-Gbps optical communications link plus a secondary high bandwidth RF comms link providing a redundant path to Earth.
- Alternatively, with today's technology (and the upcoming HPSC chipset), a more conservative and mature combination of 1 HPSC chipset plus 3 AMD processors plus a solid-state recorder based on the Roman Space Telescope design could provide 1 TFLOP of processing capacity, 20 GB of RAM, and 10 TB of solid-state storage for under 350 W average power

B. A Nebulae **Observing System in the Sky** node including processing and relay, exemplified by an orbiter at Jupiter or one of Jupiter's moons. This implementation adapts to lower available power, lower comms bandwidth, and longer comms latency. Power is the limiting factor in this design.
   - In a footprint of roughly 0.25 $m^3$, a Nebulae subsystem design in 2025 for a Jupiter-orbiting spacecraft might provide one HPSC chipset plus 2 additional processors for scientific processing and data reduction. With 1 CPU and 1 GPU, the design would provide 1 TFLOP of scientific processing capacity; the GPU performing machine vision or machine learning tasks when supporting a lander or near-real-time image processing, while the CPU supports in situ scientific analyses of any type.
   - The subsystem could have a non-volatile storage of 10 TB (persistent through power loss and restarts) and processing RAM capacity of 20 GB.
   - The subsystem would require approximately 200 W of continuous power, part of the 690 W provided by a 60-$m^2$ Solar array in Jupiter orbit.
   - An optical communications link to Earth, burst rate at 20 Gbps (not enough power is available for continuous operations).
   - A low bandwidth RF link for command and control.

C. A Storage-heavy Nebulae subsystem for a **Data Cycler** to an outer planet. For this example we opted for a larger, maximum-size one-time cycler. The design is optimized for a single high-bandwidth ingest of data at a remote area where power is constrained. Offloading data back near Earth will be easier as available power will be 5× to 20× greater (see Appendix B, Deep Space Power Generation). More specific analysis has to be performed on each possible Cycler concept to determine if the system is power-limited

or I/O limited. An RF link may create an I/O-constrained system, so we assume a medium-bandwidth optical link is used.

- For a size of 0.3 $m^3$ a Nebulae cycler design in 2030 might provide a non-volatile storage of 20 TB (persistent through power loss and restarts).
- The cycler would run a single HPSC chipset (10 GOPS) plus dedicated circuitry needed for fast I/O. Compute needs are not high for a cycler.o A "local" optical communications link with maximum data rate of 40Gbps used to offload data from remote Nebulae nodes and download the data to Earth or other nearer-Earth data relays.
- A low bandwidth RF link for command and control.
- If the optical link consumed 25 W, only for the period of the "pass," the subsystem would require approximately 125 W of power during the data writes plus 30–50 W additional for the CPU. The rough total of 200W required by the vehicle (there are no sensors) would require a 60$m^2$ solar array for a Saturn cycler or a 20-$m^2$ array for a Jupiter cycler. A 60-$m^2$ solar array may prove too expensive for a cycler.

## 5.6 Gap Assessment

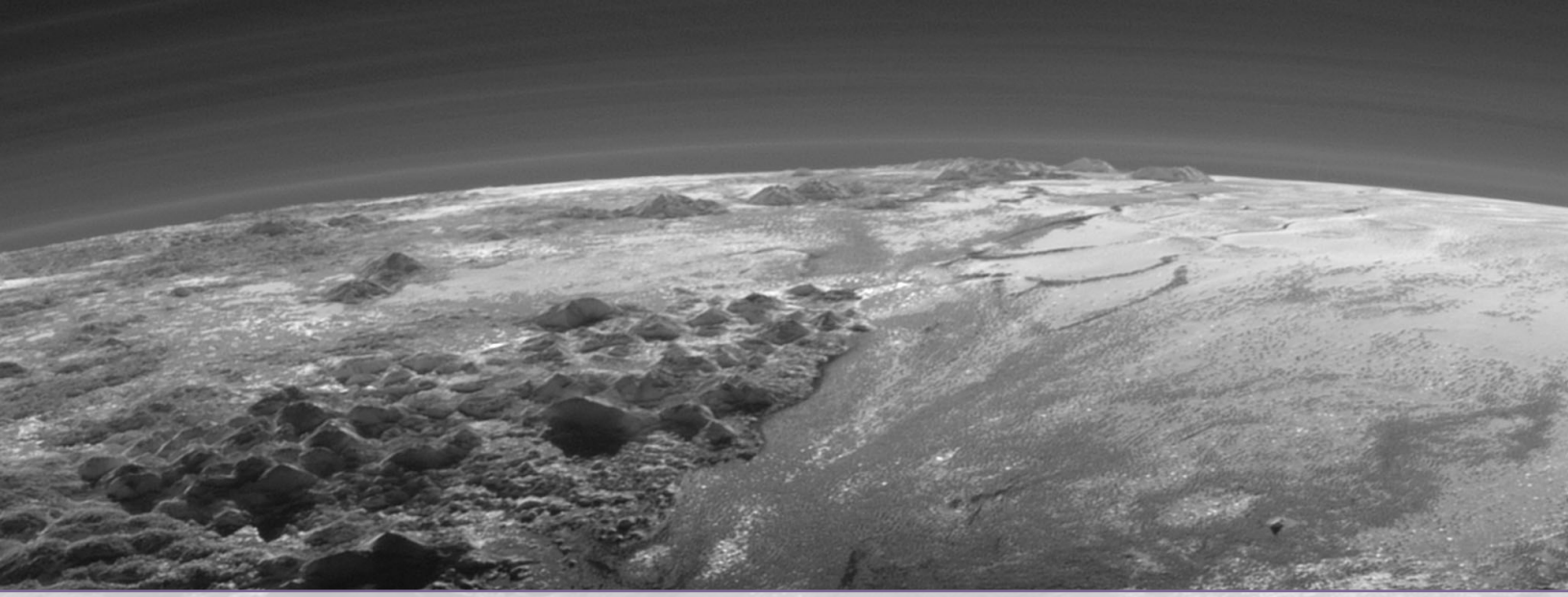
The analyses performed by the Nebulae team identified several gaps between existing capabilities and future requirements/needs.

1. **Need for higher-performance full RBHD watchdog processor:** Nebulae needs a higher performance fully radiation protected multi-core processor, such as the HPSC chipset currently in design.
   - Proceed down qualified COTS and RHBD (HPSC/GRADSOC) approaches to mitigate risk.
2. **Higher efficiency solar arrays:** RTGs are a costly power source, and dealing with radioactive materials is challenging. Achieving $\geq$40% efficiency would make Mars and Jupiter missions more flexible and would moderately improve Saturn missions' overall capabilities.
3. **Optical link capability** in the Deep Space Network, spacecraft, and possible future relay nodes: Laser/optical links will provide a significant scientific advantage due to the much greater return of data. Spacecraft will benefit from affordable 10Gbps–40Gbps optical links that consume 25–50 W.
   - We believe it is important that the DSN goes forward with its plans to add optical link capability around the globe in order to pair with deep-space optical links.

4. **Scientific methodology changes:** Currently, raw data is returned to Earth partly because the scientific community does not accept results which cannot be replicated independently by other scientists. Part of Nebulae's premise is to allow in-situ data processing and reduction. Whenever processed data is returned to Earth in lieu of raw collected signals, scientists will not be able to replicate the entire data processing chain. There must be new concepts and agreements for what is sufficient for the scientific community. This will be a major cultural and experiment design issue. At a minimum, the scientific community may require more stringent validation methods on any analytic method executed remotely, but larger cultural discussions would need to be tackled directly. If an observation cannot be repeated, how much value does it have?

5. **Resiliency as a Service:** the RaaS concepts are in the emerging state and need additional maturation before they are applied to a deep space mission. Using hypervisors to implement high-availability processor architectures and CPU voting architectures with checkpoint/restart—these been proposed but are not yet in wide use. Automated reconfiguration-on-the-fly is a core function of Earth-based public Clouds (i.e., Amazon Web Services, Microsoft Azure) but they do not offer dynamic N×M redundancy. The capabilities are maturing rapidly, driven by commercial interests. The Nebulae-based RaaS must have a much lighter footprint and only a subset of the capabilities as a full Cloud because the onboard resources are more unique and costly.
    - We recommend a separate study into Resiliency as a Service and its near-term viability.

Pragmatics desired to define a technology transition sequence that evaluates technologies on Earth, then moves to a lunar Mission, then on to a Mars mission, and finally to the outer planets. Such a progression could represent the minimum risk for each step. However, given the length of time to define missions, this progression would mean that Jupiter and Saturn missions would not see Nebulae for well over a decade. The first Nebulae mission might therefore be to Mars.



# 6. Recommendations

The KISS Nebulae workshop data can be utilized to create additional subsystem configurations supporting other mission types. As has always been the case, the final hardware and software design for a Nebulae deep-space computing and other services-enabled mission will be tailored and unique.

Nebulae nodes will leverage multi-core processor technology, emerging processors efficient at ML/DL and sensor processing, heterogeneous processor mixtures enabled by resilient hardware and software architectures, and optical communications. Some of these technologies are driven by the commercial market and evolve rapidly, while others are identified as gaps that may require nudges or significant bumps from the scientific community in order to provide enabling technology to Nebulae.

Nebulae-enabled missions should be interoperable between spacecraft launched at different times. Nebulae-enabled missions should be expected to provide scientific analysis capabilities and processing behaviors unforeseen at time of launch. This is not unprecedented. Mars missions have long carried spare communications capacity to aid in returning data from surface or orbiter missions through the Deep Space Network. NASA, ESA, and partners can and should consider on-site computing a sharable, enabling resource.

Innovative concepts like Nebulae become realized when they are developed, characterized, and socialized to the point that they become legitimate approaches to consider during mission formulation trade studies. Mission concepts emerge through the collaboration of PI scientists driven by science yield, engineers driven by risk mitigation, and programmatic considerations

driven by cost reduction. Each of these groups must be individually and jointly satisfied that a concept is sensible from their vantage point.

Scientists are currently bound to the concept of the Science Traceability Matrix (STM) as the formal process to validate that a mission addresses science needs. Missions that request additional resources beyond what the STM can validate are seen to be wasteful. Therefore, for Nebulae to become a feasible concept, identification of science use cases whose STMs would require Nebulae capabilities are mandatory. The concept of "marginal" or "spare" capacity for future mission usage, at this time, should be avoided. Rather, the focus should be placed on new or significantly enhanced science goals that justify both the general development cost and instantiation costs for any mission-specific objectives. Likewise, any discussions of or designs utilizing autonomy must seek to elevate, inform, and support scientist needs with careful attention to minimizing/characterizing inadvertently induced observation biases. Any discussion of automating science decisions, performing science onboard, or perceived "black box" approaches must be avoided at all costs. All discussions must seek to maximize what human interactions are possible, to leverage science PI feedback and control into the system ("human-on-the-loop" concepts).

Engineers are trained to "meet and not exceed requirements" and return the minimum science that satisfies the STM (binary success criteria) and the mission criteria with the lowest risk profile at an acceptable cost. In order for a mission proposal to be endorsed, engineers must be satisfied that a trade has not incurred any undue risk. For these discussions, inspiration about "upside potential" and "additional science" are anathema. It should also be accepted that utilizing new space computing systems absolutely does increase risk simply by lack of heritage, placing new concepts at a significant disadvantage. Focus should be placed on Nebulae architectures/implementations where risk can be bounded and mitigation strategies suggested. Extra emphasis should be drawn to common science mission risks that can be mitigated by Nebulae, expanding the message from "we incur more risk for more reward" to include "but we can also help reduce some current risks in unique ways."

Mission proposal reviews (and hence programmatic interests) naturally focus on mission cost: the cheaper a mission, the more missions can be funded by a given institution, program or agency. There is no doubt that space computing is expensive to develop, and significant onboard data processing / autonomous software is a new expense without strong historical precedent. A sound economic argument should be developed to demonstrate where Nebulae are cost-competitive by comparing mission formulations with and without Nebulae concepts with the same fixed science goals to make the economic benefit crystal-clear.

As all mission formulation begins with the science PI and science goal, digesting Nebulae offerings into an easily-to-assimilate menu of science-facing capabilities becomes the critical interface for impactful infusion into the trade-space discussion. These capabilities should

not be phrased in architectural, data processing, data science, or engineering jargon but rather made clear using language familiar to the science PI. The effective benefit of each capability should be illuminated by actual science use cases drawn from past and current missions (hence, the importance of the retrospective studies), with the costs in terms of risk and spacecraft resources made clear, and the alteration of the science data processing pipeline/concept of operations fully explained from the perspective of "a day in the life of a mission scientist." The emphasis should be placed on full transparency of positives and negatives, providing both key drivers for both applicability and non-applicability.

In short,

- Leverage the communications relay precedent to include sharing of compute / storage resources, possibly including in-stream computing of data being relayed or cached. A relay node already requires significant compute, why not have it perform data prioritization or summarization as well?
- Engage NASA mission reviews to ask "Have you analyzed the potential science gain for a modest increase in onboard compute/memory?"
- Start "Opportunity Cost" tracking for ongoing missions and the hardware/software that could have captured what was lost.
- Foster a discussion across the scientific community to discuss under what conditions research and conclusions drawn from non-repeatable analysis and experiments can be utilized and accepted by the scientific community, shared and amortized across multiple missions. Contrast that to a rigid STM calculation that treats each mission independently.
- Perform an economic analysis comparing scientific data collection and analysis with and without Nebulae capabilities.
- Define quantified capabilities benefits for future deep space missions, illuminating the science benefits of Nebulae capabilities to specific missions.

In regards to multi-spacecraft campaigns such as those that have been deployed to Mars and Earth,

- Incentivize mission designers to consider how their mission could benefit from utilizing resources from existing spacecraft, such as comms relay or onboard data, and how future missions might benefit from theirs.

Acknowledging the unique resource situation of Earth observing, and the evolving considerations of New Space,

- In the Earth System Science arena, seek a win-win partnering concept between government and industry which acknowledges and provides room for different objectives (business model vs. science understanding) and draws on the respective strengths of each.

# Appendix A: Market Trends for Spacecraft

## A.1 Deep Space Communication

Deep space missions to date have always been able to collect more raw sensor data than can be returned to Earth because of communications systems constraints. Studies have indicated that NASA's deep space science missions (assuming similar operating paradigms) will tend to return an order of magnitude more data with each coming decade.[1] This presents a challenge to the designers of these missions. NASA expects to keep up with this predicted demand for the next couple decades at least. Consider the NASA Deep Space Network (DSN) chart in Figure A.1, which shows planned communications capabilities from various places in the Solar System to Earth.

The DSN supports over 30 missions concurrently, using fewer than 30 dishes. Because of the geometry of the rotating Earth and orientation to each distant mission, no mission has access to the full terminal bandwidth all the time. Additional dishes extend the overall DSN capacity and number of missions supported.

Since communications performance is proportional to the inverse of the square of the distance between transmitter and receiver, it is particularly hard to move data across deep space distances, as indicated by lower actual and projected data rates. Even with an order of magnitude improvement every decade, communications will not catch up with the ability of sensors to collect data.

[1] Deutsch, L. J., Townes, S. A., Liebrecht, P. E., Vrotsos, P. A., & Cornwell, D. M. (2016). Deep space network: The next 50 years. In 14th International Conference on Space Operations (p. 2373).

It is assumed in these predictions that the spacecraft in question has a similar capability (antenna size, available transmitter power, pointing accuracy, etc.) to the Mars Reconnaissance Orbiter (MRO). Figure A.1 predicts the approximate data rate on the downlink from these spacecraft. Uplink (moving data from Earth to the spacecraft) has not been a bottleneck up to this point in time, as it has typically consisted of a few sparse "commands," some sequences (lists of commands), and the occasional software patch. It should be noted that radio links from deep space with data rates beyond ~3 Gbps will be unlikely due to international radio spectrum constraints. This does not apply to optical communications, which is one of the reasons NASA will likely move to this in the future.

Data rates between Earth-orbiting spacecraft at geosynchronous earth orbit (GEO) and below are already challenged by the spectrum limitations. Overall space-to-ground communications capacity is increasing exponentially over time, so we can assume Gbps links are available for any spacecraft data system that has the ability to handle Gbps.

Likewise, crosslinks—links between spacecraft that are in orbit about the same planet—can already support tens of Gbps. This has been demonstrated in Earth orbit. NASA's Tactical Data Relay System (TDRS) satellites allow the International Space Station, the Space Shuttle, and other satellites to have simpler on-board communications systems and send data to TDRS; then, TDRS's larger and higher-powered system downlinks through the atmosphere.

| | Distance (AU) | Today (Mbps) | 2025 (Mbps) | | 2035 (Mbps) |
|---|---|---|---|---|---|
| DSN Configuration | | 34m X-band | 34m Ka-band | 34m Ka-band | 8m Optical |
| Spacecraft Configuration | | 3m antenna 100 W transmitter 1/6 Turbo code | 3m antenna 100 W transmitter 1/2 LDPC code | 3m antenna 180 W transmitter 1/2 LDPC code | 30cm antenna DSOC-like power and coding |
| Venus (Closest) | 0.3 | 80.0 | 320 | 576 | 3200 |
| Venus (Farthest) | 2.4 | 1.3 | 5 | 9 | 50 |
| Mars (Closest) | 0.6 | 20 | 80 | 144 | 800 |
| Mars (Farthest) | 2.6 | 1.1 | 4 | 7.67 | 43 |
| Jupiter * | 5.4 | 0.247 | 0.99 | 1.78 | 4.94 |
| Saturn | 10.1 | 0.071 | 0.28 | 0.51 | 1.41 |
| Uranus | 19 | 0.020 | 0.08 | 0.14 | 0.40 |
| Neptune | 30.3 | 0.008 | 0.03 | 0.06 | 0.16 |

Figure A.1: Maximum per-link (per connection) communications capacity for the DSN as of 2020. DSN antennas must be facing the spacecraft in order to receive data.

Crosslinks and relay systems like TDRS also provide downlink of data from spacecraft which do not have direct line-of-sight access to the DSN terminals, and can provide additional spectrum to utilize for the final miles of downlinks.

Crosslinks can be much higher in performance, particularly when the spacecraft are orbiting other planets that are not constrained by the spectrum law. 10-Gb to 40-Gb low-cost optical crosslinks will be available in volume by the mid-2020s, driven by the market demands of large commercial constellations and government low-Earth orbit (LEO) constellations such as the Space Development Agency's National Defense Space Architecture, which will have hundreds of satellites in orbit by 2030, each with 1 to 4 optical crosslinks that initially operate at a range of 1,500 miles, and DARPA's Blackjack program.

All of these example calculations assume MRO-like spacecraft systems. With the trend toward smaller spacecraft, this might be a poor assumption—smaller spacecraft generally have lower communications capacities. Spacecraft farther from Earth have additional communications limitations. Also, it is entirely possible that downlink data rates might increase faster than 10× in a given decade. The reasons for this are explained in the "Science" section.

Hence, even with the most optimistic growth in communications systems capabilities for returning data, NASA needs to be considering other alternatives for increasing the return of information. Notice again the difference in terminology.

The ITU is currently working on rules for outer body spectrum as a resource and for in-site use on celestial bodies. For now, given the sparsity of deep space assets, we assume the spectrum is only regulated (loosely) on/near the Moon and Mars.

## A.2 Processing Capacity

The commercial market drives innovation in processors, and those processors are always released first for the major market—which means they are designed to survive the mild radiation experienced on Earth. Companies sell thousands or millions or even billions of each commercial processor. The Radiation-hardened by design (RHBD) chip market is financed solely by NASA, NOAA, and other space agencies; defense and intelligence agencies; and the commercial marketplace for satellite communications and earth sensing—which together drive volumes in the tens or hundreds for each processor, rarely reaching 1,000.

The relatively tiny RHBD market drives a large cost to obtain the same performance a commercial processor provides, because the non-recurring cost is amortized across a much smaller production run. Therefore, while the Nebulae core C&DH processor and spacecraft "watchdog" functions will always reside on a RHBD processor, a fully RHBD Nebulae node will always be expensive and likely unaffordable. We looked to ways to increase compute

power by trading inherent radiation performance for processing capacity, and addressing radiation impacts in other manners.

### A.2.1 Processor Capacity: CPUs

Radiation-hardened by design (RHBD) CPUs are approximately 2 to 3 technology generations and 100× processing capacity behind commercial state-of-the-art CPUs. Alternatively, one could view the RHBD CPUs as 10 to 15 years lagging commercial CPUs. This will always be the case because the commercial market drives innovation because of higher production run quantities and much lower demands on radiation tolerance.

Figure A.2 shows GOPS (billions of operations per second) over time. Compare the orange and yellow lines to the blue and grey lines; the difference may be compared by years to equivalent processing capacity in space (horizontally) and by processor generations if one compares a RHBD processor to the commercial equivalent. There are many more commercial processor families than are represented by the orange and yellow lines, however there are not many more RHBD processor families than shown.

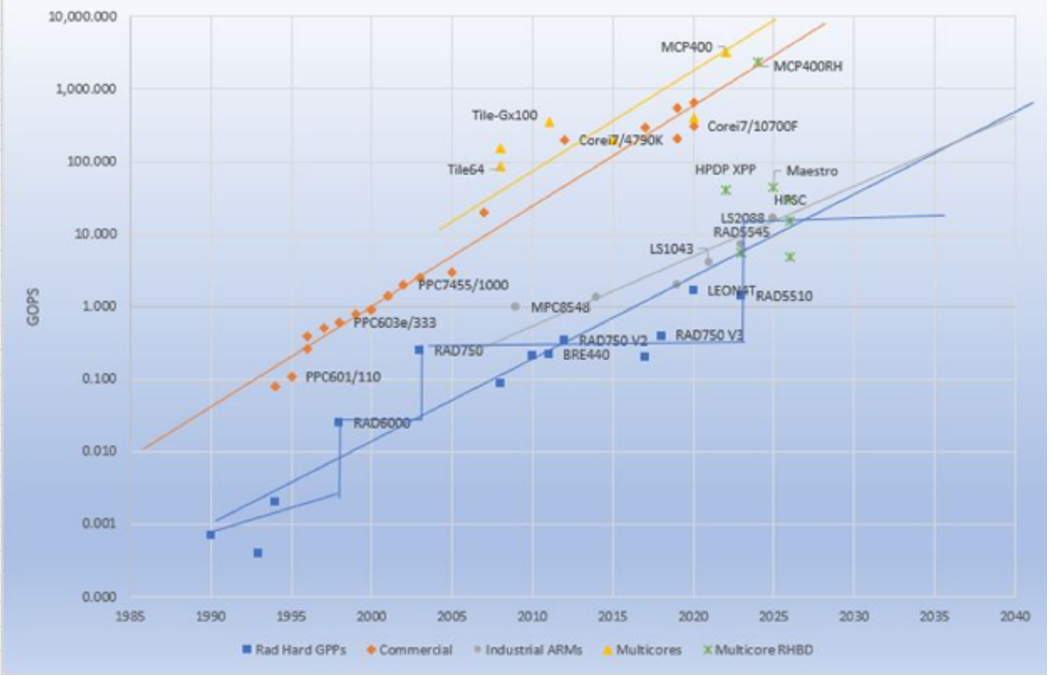



Figure A.2: Processor capacity growth over time.

A key recent technological advance that enables Nebulae is the emergence of multi-core CPUs. Looking again at Figure A.2, notice how the gap between commercial and RHBD processors has widened since the release of the BAE RAD750 processor in 2003. RHBD single-core processor capacity has stagnated. However, the emergence of multi-core RHBD

processors (green Xs) increases capacity by 1 to 2 orders of magnitude and reduces the gap to commercial processors more towards historical norms.

The most important impact of multi-core CPUs is that processing capacity is moving on board in the commercial and government markets. As mentioned before, the root and core of the vehicle's control system—the command and data handling (C&DH) subsystem—cannot fail. C&DH software is rigorously tested, is rarely altered after launch, is partially or fully written in hard-real-time methods, and is isolated from sensor processing and other software. C&DH control software was allocated one entire single core processor. The advent of multi-core processors (4, 8, and soon 16 cores per processor) means that the C&DH can be protected and isolated to a core, and suddenly there are 3, 7, or even 15 additional cores of capacity available for the mission, essentially free from a SWaP and cost perspective. This motivated the commercial satellite market to create layered control and application architectures that can take advantage of the new capacity. These new architectures directly enable Nebula's multi-processor capabilities and in-flight mission processing upgrade capabilities.

CPU chip power consumption has increased recently as performance increases. Some new processors draw in excess of 100 W per chip, limiting their utility in power-constrained deep space missions.

### A.2.2 Processor Capacities: GPUs

The layered software architectures that evolved to support multi-core processors are also extensible to support multiple heterogeneous processors. This boon enables us to consider other recent advances in processor design, most importantly the Graphics Processing Unit (GPU) and recent GPU evolution, which supports machine learning (ML) and deep learning (DL) artificial intelligence software.

Graphics processors, sometimes referred to as GPPs or GPUs, were created to efficiently process real-time visualizations required for video games and for high-performance renderings for complex images. GPUs are dramatically more efficient than general purpose CPUs for certain types of algorithms, including graphics processing for machine vision, signal processing, machine learning, and matrix manipulation. GPUs can provide an order of magnitude of raw performance benefit in these areas over a single core CPU—note the figure below. As the single-core CPU capacity growth slowed, GPUs continued their rise. (See Figure A.3)

The self-driving car (autonomous vehicle) market is driving rapid miniaturization of GPUs as well as specific hardware changes designed to support low-cost self-driving—which, coincidentally, is exactly in line with certain deep-space sensor processing needs and autonomous flight control using LIDAR returns.

Graphics processors currently are the processors of choice in the commercial market for machine learning (ML) and deep learning (DL) systems. The efficient matrix manipulations

and parallelized architectures provide up to 50 times improvement in execution speed over general purpose CPUs. These processors are desirable for autonomous lander flight processing.

There is no history of RHBD graphics processors tuned for ML/DL execution, and as of 2020 there are no plans by industry leaders to produce an RHBD GPU. As GPUs are environmentally hardened for the automobile industry and self-driving cars, we may be able to select GPUs from those production runs that are radiation-tolerant enough to be a high-performance sensor processor.

A 2020 example is the Nvidia Jetson Xavier processor, which benchmarks at 21 TOPS (trillion operations per second) on 8-bit integer calculations while consuming 15 W of power. One of the targeted purposes of this GPU is to process 1080p video.

GPUs can consume even more power than CPUs; the most recently announced ML GPUs consume over 200 W per processor, too much for any missions beyond Mars.

### A.2.3 Technology Evolution and Maturation

As of 2020, GPUs are used as high-performance mission processors, not as flight computers. The general nature of the software used for guidance, navigation, flight control, and error handling favors a general purpose CPU.

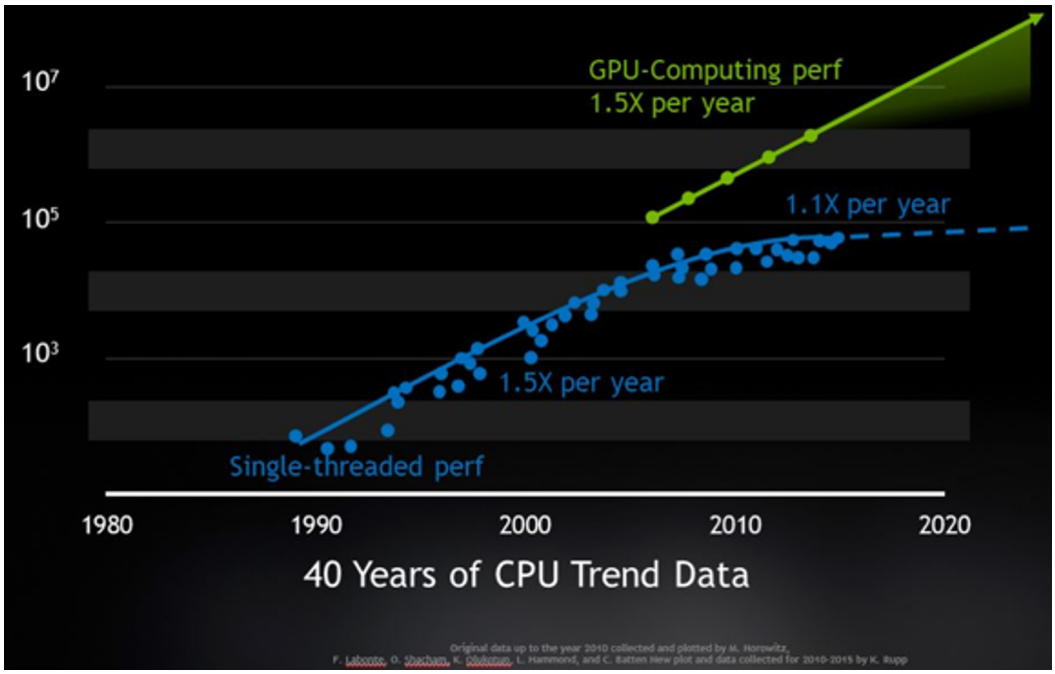



Figure A.3: General purpose CPU compared to Graphics Processing Units (GPU).

There are emerging low-power GPUs designed to optimize machine learning code, including Google and Nvidia tensor processors. Even lower-power processors do exist which have proven to be radiation-tolerant, such as Qualcomm's Snapdragon cell phone processor. The available compute power will be significantly expanded if small commercial CPUs and GPUs turn out to be highly radiation tolerant without any investment from the space community.

The marketplace for GPUs is highly dynamic. Emerging 5G cellular networks are funding massive changes in low-power high-performance "industrially rugged" processors. There is a merging of GPPs into FPGAs and Arm (advanced RISC machine) processors, which will be accelerated by Nvidia's purchase of the Arm company. The supporting software environments are maturing rapidly in parallel, and we can expect robust development and execution environments for complex software on small, low-power processors of all types.

### A.2.4 Silicon Feature Size

As shown in Figure A.4 below, manufacturers have used ever-shrinking mask sizes to achieve greater gate density, computational power, and capacity. Commercial foundries operate now at 7 nm with 3 nm capabilities soon to come on line and immediately operate at full capacity for Apple's iPhone and other consumer products. However, that cannot continue—the line can never reach zero and quantum effects become significant near 1 nm. Similarly, clock speeds cannot increase indefinitely. This is part of the motivation to move to multiple cores and chiplets.

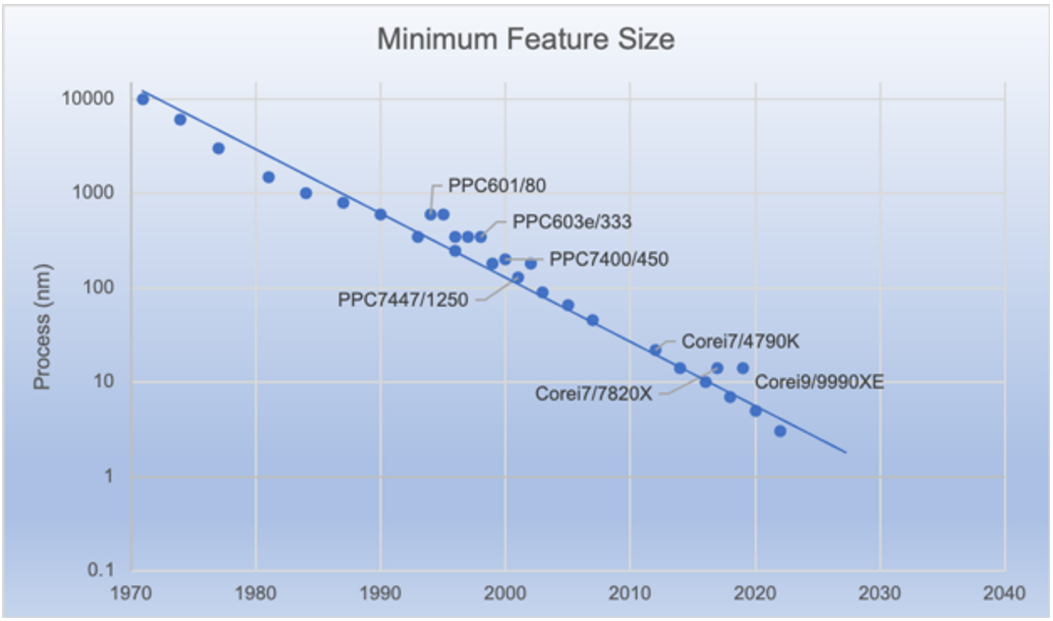



**Figure A.4:** Processor feature size reduction over time. Silicon mask sizes are reaching a minimum due to electron flow requirements and quantum effects.

## A.3 Data Storage Capacity

Radiation-hardened memory (non-volatile storage) lags 10–20 years behind commercial technology, equivalent to about 2 generations of technology or 2–3 orders of magnitude of capacity. Plots of commercial and RHBD memory are shown in Figure A.5—the yellow points being non-RHBD and blue points showing RHBD chips. The commercial trend was linear while mask size decreased and recently is more scattered as other techniques are used to increase capacity. RHBD trends are harder to identify; it is possible that the gap is widening and that RHBD lags by more than 2 orders of magnitude.

RHBD hybrids are approaching 1 GB per module. Using 1 GB to represent 2020 state of the art, we assume 10 GB modules will be available in 2030 for our Nebulae "what if" scenarios.

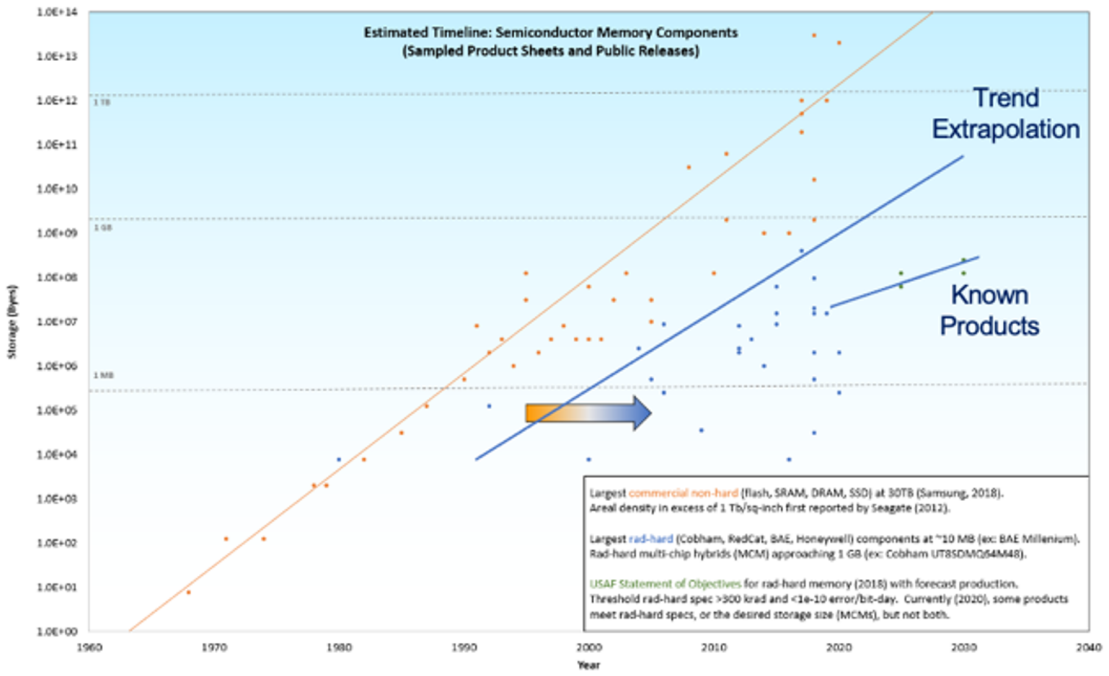
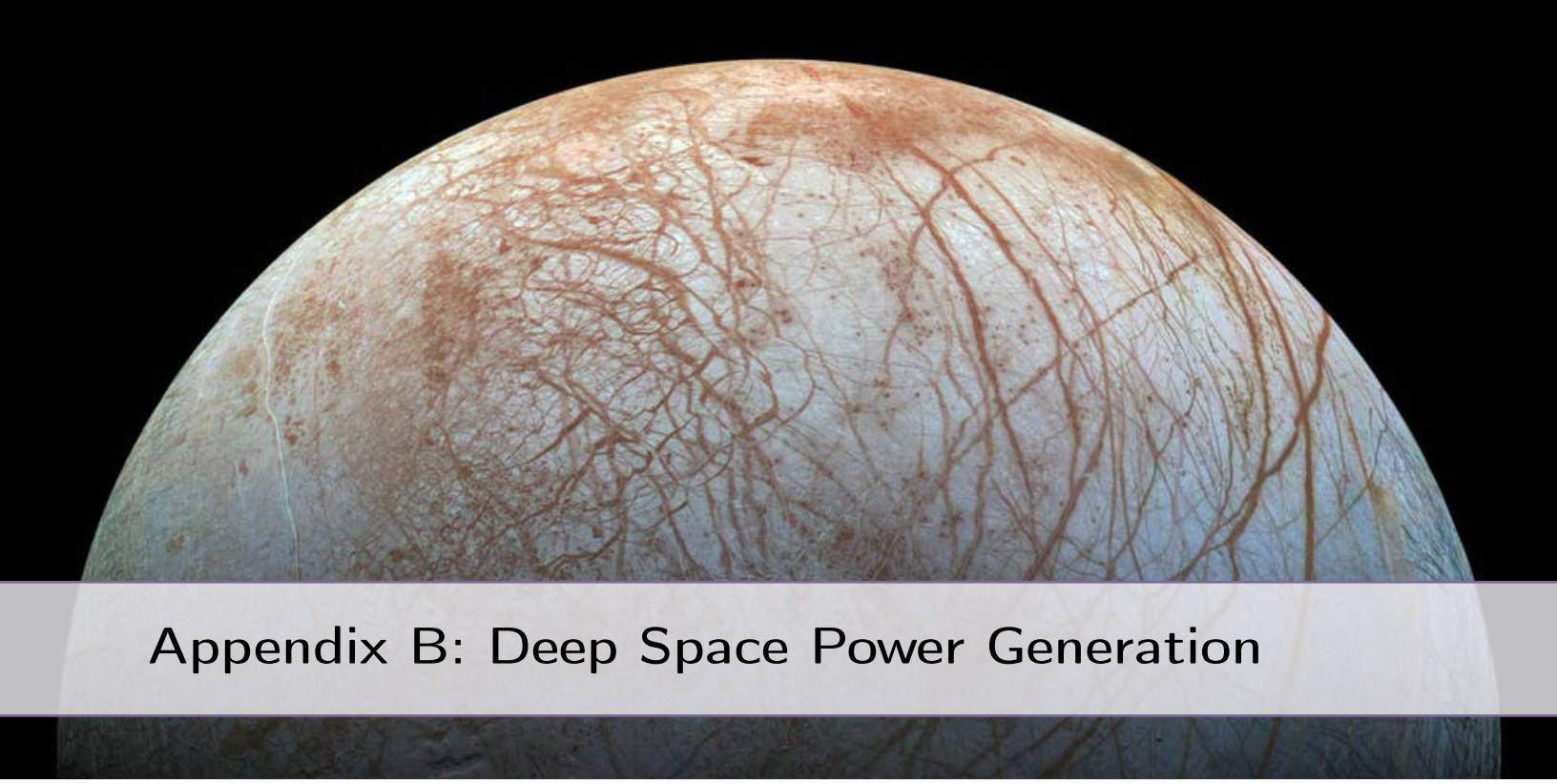



Figure A.5: Semiconductor memory capacity growth over time.



# Appendix B: Deep Space Power Generation

Power generation is critical for spacecraft. Communications data rates are directly proportional to the electrical power applied. Sensors and processing units are the other consumers of power. Power limitations often cap the communications to Earth, which in turn limits total mission science delivered.

Spacecraft electrical power is provided in 3 ways: via solar arrays, radioactive thermoelectric generators (RTGs), or small fission reactors. Available power in 2025 through 2035 is extrapolated for solar arrays and radioisotopic power generators. The United States only launched one reactor, back in 1961, and is unlikely to do so to power a spacecraft. A potential reactor for a Mars base will be too heavy for reuse on a spacecraft.

For solar arrays, we assume 3% power efficiency improvements in solar arrays every 10 years based on 40+ years of commercial solar array evolution, as shown in the Figure B.1 chart from NREL. Ground-based solar arrays for commercial use conservatively average 4% gain every decade, not including rapid startup efforts to reach 20%. Some spacecraft manufacturers advise that 4% is aggressive and that the rate of improvement is slowing down for space-based arrays, hence our selection of 3% per decade.

Figure B.2 shows solar array power generation capacity normalized to a 60-$m^2$ surface area, using Juno and a commercial satellite launched in 2019 as baseline data, extrapolated through 2035. The Juno mission used a 60-$m^2$ array, which is used as the reference point.

There have been 7 generations of **Radioactive Thermoelectric Generators** (RTGs) since 1961. No significant increase in capacity or capability is foreseen in the near future. The Advanced Stirling Radioisotopic Generator (ASRG) could have been ready in 2026; however,

development and fielding was halted within the last several years. The ASRG would provide 6 times the efficiency of previous RTGs and simplify thermal management. RTGs are predictable, reliable, inefficient, expensive, and require nuclear material from the Department of Energy. RTGs are often the recommendation for far space missions and large landers/rovers. The U.S. Radioisotope Power Systems Program is managed under SMD/Planetary Science at the Glenn Research Center.

**Fission reactors** are not a practical option for U.S. Nebulae systems. While the USSR/Russia has launched dozens of fission reactors on spacecraft, the U.S. has launched only one, in 1961. Reactors are being considered now for off-planet "permanent" 1.0-kW to 10-kW power generation in a Mars colony, under the KiloPower and Kilopower Reactor Using Stirling Technology (KRUSTY) initiatives. The Kilopower project is under TDM and is now called Fission Surface Power managed out of the Glenn Research Center. Reactors are too heavy for nearly all spacecraft and have very large heat dissipation requirements.

From a myopic power-conservation perspective, the most effective means of transmitting lander sensor data to Earth is sometimes to use an orbiter as a relay. The orbiter can carry a larger unimpeded solar array to generate more power and can carry a larger reflector, which together drive greater data rates back to Earth than a lander. Relay spacecraft can also be oriented toward Earth a larger percentage of time than a lander. However, relays add considerable complexity and cost relative to a single vehicle mission.

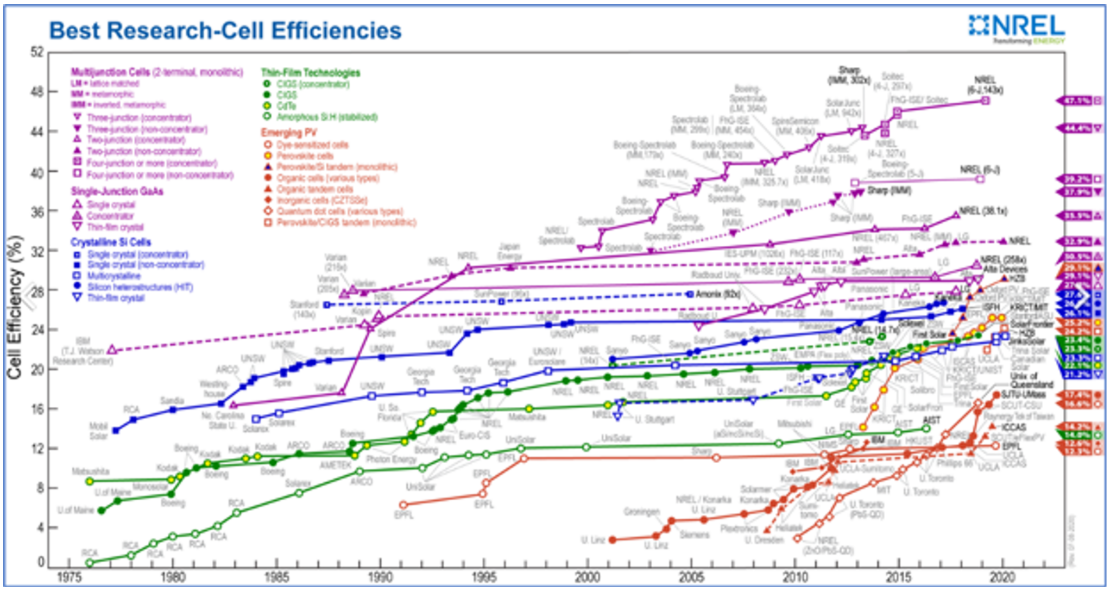



Figure B.1: Solar array panel energy conversion efficiency (Credit: NREL).

## B.1 Power Generation Summary

Fission reactors are not practical for spacecraft, leaving the choice between RTGs and solar arrays. Solar arrays are more desirable for many reasons—ease of manufacturing, launch safety, more plentiful resources—but an improvement in solar array efficiency to 40% or

| Solar Arrays | Distance (AU) | Available solar energy (kW/m$^2$) | 1990 (kW) | 2011 (kW) | 2020 (kW) | 2025 (kW) | 2030 (kW) | 2035 (kW) |
|---|---|---|---|---|---|---|---|---|
| Efficiency | | | 15% | 17% | 25% | 27% | 28% | 30% |
| Mercury | 0.4 | 8.95 | 69.0 | 78.2 | 115.1 | 124.3 | 128.9 | 138.1 |
| Venus | 0.7 | 2.63 | 20.3 | 23.0 | 33.8 | 36.5 | 37.8 | 40.5 |
| Earth | 1.0 | 1.36 | **10.5** | **11.9** | 17.5 | 18.9 | 19.6 | 21.0 |
| Mars | 1.5 | 0.59 | 4.5 | 5.2 | 7.6 | 8.2 | 8.5 | 9.1 |
| Jupiter | 5.2 | 0.050 | 0.39 | **0.44** | 0.65 | 0.70 | 0.72 | 0.78 |
| Saturn | 9.5 | 0.015 | 0.12 | **0.13** | 0.19 | 0.21 | 0.22 | 0.23 |
| Uranus | 19.2 | 0.004 | 0.03 | 0.03 | 0.05 | 0.05 | 0.05 | 0.06 |
| Neptune | 30.1 | 0.002 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 |
| Technology | | | 60 m$^2$ GaAs | 60 m$^2$ Triple junction GaAr | 60 m$^2$ 3rd Gen Triple junction GaAr | | | |

Figure B.2: 60-m$^2$ solar array power generation capacity at each planet.

| | TRL9 Year | Power (W) | Power Density (W/kg) | Heat Dissipation (W) | Efficiency | Missions | Recurring Cost |
|---|---|---|---|---|---|---|---|
| GPHS-RTG | 1980 | 290 initial, 213 EOL | 5.3 | 4400 | 5% | Cassini x3, New Horizons, Voyager x2, | $118M |
| MMRTG | 2012 | 110 initial, 55 EOL | 2.6 | 1880 | 6% | Mars Science Lab, Curiosity, Mars 2020, | $109M |
| ASRG | 2026 | 135 | 4.9 | 250 | 29% | Uses RTGs | |
| NexGen RTG | 2030? | 500 initial, 362 EOL | 6.0 | | | | |

Figure B.3: RTG power generation capacity.

greater will enable greater utility on Jovian missions and deeper into the solar system. RTGs are currently the only viable power source for missions beyond Saturn.

The selection of power source is different for landers, whether they drive or fly. Landers cannot carry a large solar array like the 60-$m^2$ array flown on Juno. Landers only collect solar energy when facing the Sun and available solar energy is reduced by atmospheric interference. Landers using solar arrays therefore generally have significant constraints on duty cycle and available power because of the limitations on collecting energy. Additionally, only a subset of landers are large enough to carry an RTG's weight.

In 2026, the crossover distance in available power between a 60-$m^2$ solar arrays and a single 135-W ASRG RTG source will be 9.4 au, slightly before Saturn at 9.5 au. Today the cutover point for a single 110-V MMRTG source is at 10.4 au, slightly beyond Saturn. Those, however, are initial values and the RTGs decay over time. The array power is reduced by the square of the distance from the Sun but is generally constant for the lifetime of the spacecraft.

Beyond Saturn, more power is available via an RTG. RTGs will continue to be the preferred power source beyond Saturn, and because that power is independent of vehicle orientation or line of sight to the Sun, may remain the best choice for missions to Saturn and its moons.

For any missions inside of the Jupiter orbit, solar arrays will remain more capable and practical.

# Appendix C: Resiliency as a Service

There are two primary types of applications that will run on Nebula: real-time analysis applications and science data processing applications:

$R$: Real-time analysis applications, such as change detection, need to be guaranteed sufficient compute resources during their duty cycle.

$S$: Science data processing applications include two subclasses:

- Production of standard data products, e.g., for different levels, and
- Data analysis used to answer research queries.

For simplicity of analysis here, we assume that an application in $S$ runs on the same-sized dataset and requires the same compute resources each time it runs. In the absence of events that preempt standard processing, such as a detected change, mission goals should allow us to characterize applications from classes $R$ and $S$ and their computational requirements for some period of time $T$:

$R$: Application $r$ in $R$ requires $op\_rate(r)$ operations (e.g., flop) per second for the fraction of $T$, $duty\_fraction(r, T)$, that it is running.

$S$: Application $s$ in $S$ requires $dataset\_ops(s)$ operations to process run once and will need to run $dataset\_rate(s, T)$ times during period $T$.

While the performance characterization functions $cycles\_rate()$ and $dataset\_cycles()$ are target machine dependent, there are likely only small differences across different instantiations of the same computer architecture. Furthermore, despite the differences across architectures, an operation rate or count on any could be used to make first order estimates of compute requirements.

## C.1 Sizing the Compute Resource: Capability and Capacity

With a characterization of the applications to run, their frequency of use, and the resources they require to run, we can make an initial estimate of the compute requirements needed for a specific Nebula-enabled mission by looking at the total requirements for $capability$ (the maximum compute requirement at any one time in operations/second) and for $capacity$ (total operations available over some period of time). Note that these are analogous to power and energy in measuring electricity. To determine the system's capability requirement, we need to examine the overlap in the running of the real-time applications. In particular, where $Running(t)$ is the set of applications in $R$ running at time $t$, we can define the minimum capability required to run the applications in $R$ as:

$$minimum\_capability\_op\_rate_R = \max_{t \in T} \sum_{r \in Running(t)} op\_rate(r)$$

With this definition, $capability_R$ is the operation rate that the Nebulae system must be able to achieve to run the realtime applications alone, given that their overlap is described by $Running(t)$.

In addition to the capability constraint imposed by the realtime applications, the system must also be able to meet the capacity requirements of all applications. We can calculate the number of operations required during time $T$ to run the applications in $R$ and $S$ as:

$$capacity\_R(T) = seconds(T) \times \sum_{r \in R} op\_rate(r) \times duty\_function(r, T)$$

$$capacity\_S(T) = \sum_{s \in S} dataset\_ops(s) \times dataset\_rate(s, T)$$

where $seconds(T)$ is the number of seconds in time period $T$. Then, the sum $capacity_R(T) + capacity_S(T)$ is the minimum number of ops required in time period $T$ to process the workload in $R$ and $S$. The minimum operation rate (in ops/second) to support that workload is thus:

$$minimum\_capacity\_op\_rate_{R \cup S} = \frac{capacity_R(T) + capacity_S(T)}{seconds(T)}$$

The starting point for designing the Nebulae compute resource would be:

$$baseline\_op\_rate_{R \cup S} = max[minimum\_capability\_op\_rate_R, minimum\_capacity\_op\_rate_{R \cup S}]$$

In addition to capability and capacity requirements, the Nebulae system should consider that cycles will be "lost" for a variety of reasons, for example:

- Resource monitoring and allocation will require cycles.
- Resources cannot be scheduled so that they are 100% utilized.
- Some computations will fail and need to be redone or restarted from a checkpoint.
- The system will lose capacity over time due to component failures.
- Redundancy and resiliency implementations will add overhead.

## C.2 Historical Comparisons

Previous scientific applications have demanded computational capacities on the order of 1–10 GOPS (billion operations per second) for autonomous mission planning; 10–50 GOPS for fast traverse and landing, advanced state-of-health monitoring, and space weather processing; and from 50 GOPS to hundreds of GOPS for radar science and hyperspectral image processing and robust scientific analyses.

Digital signal processors historically are utilized for efficient matrix math, linear algebra, and FFTs. General purpose processors are typically used for control & decision processing, searches, and general math. Graphics processors are emerging for neural network math and sparse-matrix processing. The equations above are valid in all cases, but use very different capacities for each processor type.

## C.3 Other Considerations

The equations of the previous section assume that the operations count requirements of applications are the tall pole in resource allocation considerations. In fact, some applications may have very large memory footprints that, with some solution architectures, will force the use of extra CPUs in order for those memory requirements to be met.

Single-precision math requires half the memory resources of double-precision math, and perhaps half of the computing resources. To the extent possible, the actual scientific algorithms intended for the Nebulae processor should be characterized and taken into account when selecting the processors.

The compute resource on a Nebula-enabled mission is not standalone. It competes for power with other systems, such as communications. It may be throttled back in times of very high power demands; it may be able to use surplus power at other times. In a sense, power is a resource to be scheduled, just like communication time, instrument time, archive time, or compute time.

## C.4 Benchmarking Processors

As processors evolved and incorporated significant added complexity, comparing processors by benchmarking has become more complicated. Processors have millions more gates than several decades ago. These gates implement special-purpose circuits that may be unique to one processor. Previously, scientists would write their processing algorithms and tune them for a general purpose processor. Now, there may be processors with unique enhancements or unique architectures that can execute certain processing algorithms natively one or even two orders of magnitude faster.

Benchmarking isn't just counting MIPS or FLOPS these days. The engineer has to determine what is the appropriate benchmark for classes of software and algorithms that will be running on board? This will significantly influence processor selection because different processors are tuned to optimize different benchmarks.

Processor benchmarks available include but are not limited to simple counts of speed:

- Instructions per second / operations per second (MIPS, GOPS, TOPS)
- Floating point operations per second (FLOPS, GFLOPS, TeraFLOPS),

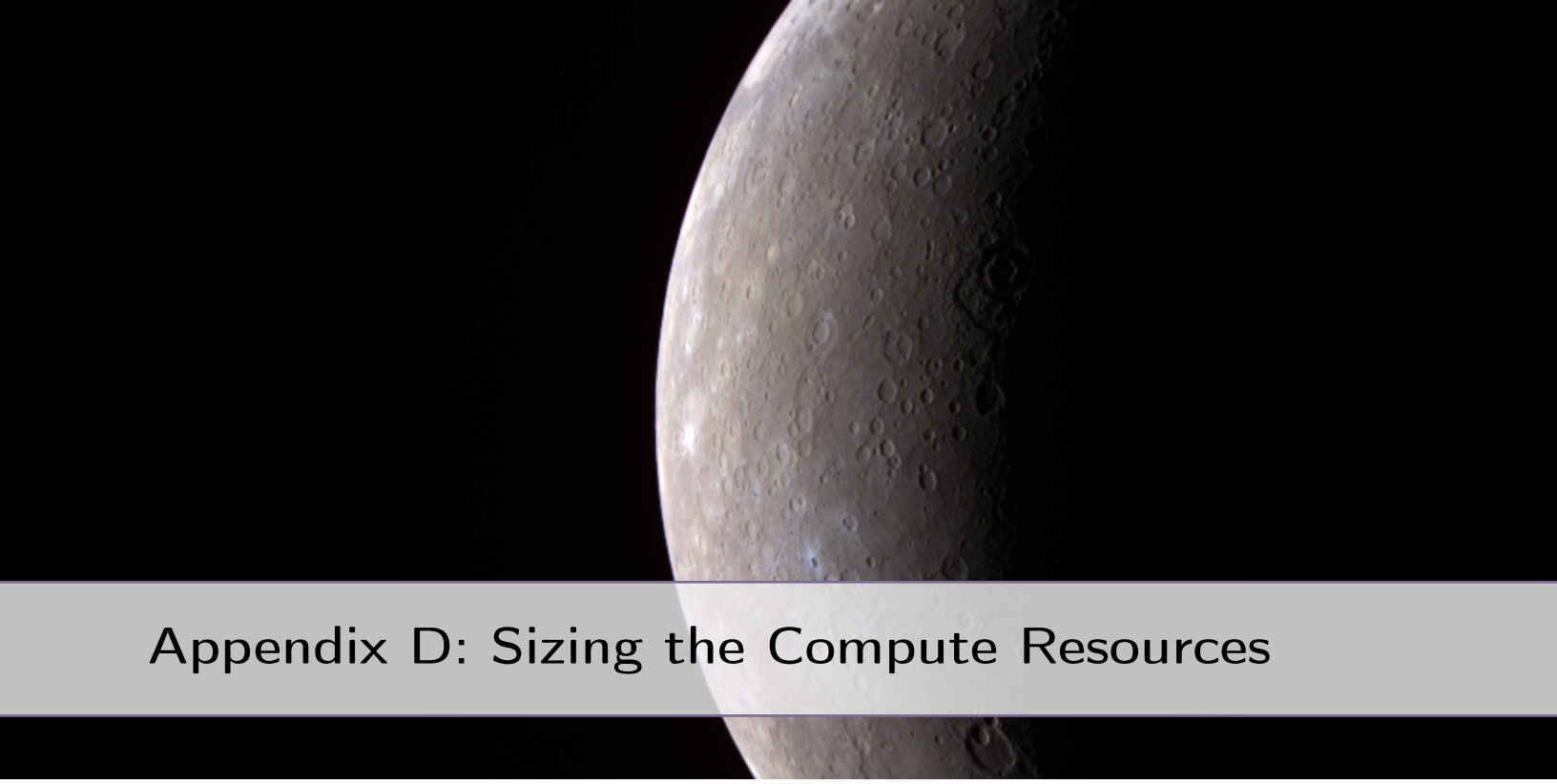
as well as various standard mathematical functions:

- Floating point measurements at various depths (FP16, FP32, FP64)
- Integer calculation measurements at various depths (INT4, INT8, INT16)
- Tensor calculations
- Chi Sum, MXTEL, BE128.

Additional complexity in multi-core processors means that different benchmarks yield different results. For example, the Nvidia A100 GPU measures at peak 312 TFLOPS when performing a

32-bit floating point ML training benchmark, while measuring 19.5 TFLOPS when performing a more traditional 64-bit floating point high performance computing benchmark. Various cores within the processor vary by up to 100× in their performance as they are optimized for different purposes. (Note: the A100 consumes up to 400 W of power and is an enterprise-class processor not suitable for Nebulae; however it is a good example of 2020 state of the art complexity in benchmarking.)

**Characterizing the mission software that will be executing on the spacecraft is therefore critically important to benchmarking and selecting the processors.**



# Appendix D: Sizing the Compute Resources

The resiliency of a system is measured using a combination of hardware characteristics derived from wiring diagrams and component failure rates together with software resiliency characteristics such as implementations of redundancy, automated failover, and recovery. Higher resiliency nearly always costs more than lower resiliency in some manner: hot spares are essentially wasted capacity; RAID solutions consume more storage space per bit of data than non-RAID; watchdog processors and software consume power and space without providing direct mission value.

Traditionally, this is a static calculation. If the software characteristics are altered—let's say the data in storage was implemented in a RAID5 configuration—then the resiliency of the system is altered. **Resiliency as a Service** means the processing subsystem's resiliency attributes are scalable, adaptable, and dynamically adjustable in order to make science data processing and storage fault tolerant.

To dynamically change the Resiliency posture:

- The N×M redundancy scheme of multiple CPUs can be changed on the fly
- Hypervisors provide restarting of non-mission-critical processors
- The RAID posture can be altered, and data stores can be configured independently.
- Layered software abstracts the CPUs from the applications, allowing for movement of programs to different resources for efficiency and resiliency.

- Software provides checkpoint/rollback/restart functions for failure/fault recovery at any middle layer of software
- Software can reset much of the hardware—as a goal, all of it; in particular, the core software on the core computers can reset/restart any of the additional processors and sensor processors.
- Software provides the ability to re-process scientific data if a failure occurred during processing.
- Cloud capabilities support processor-agnostic software, uploadable new algorithms, and software-based resilient behaviors

RaaS requires multiple processors and cross-connected storage and I/O, plus a well layered software architecture, plus some additional emerging software technology.

To take advantage of multiple CPUs and dynamic RAIDs, the second key characteristic is onboard reconfigurable software. Nebulae will be able to upload software from researchers on Earth throughout the mission lifetime. The uploads will be carefully planned and managed to protect mission data; however, the concepts are already very mature in commercial clouds. Commercial clouds present an extremely reliable infrastructure at the operating system and storage unit level—appearing essentially perfectly reliable to the consumer who needs processing and storage. The same layered, dynamic software approach will be employed in deep space to allow the scientific algorithms and sensor processors to assume a near-perfect environment, while the software automatically identifies system problems, masks impacts from the mission, and resets, restarts, and reconfigures the processors and software processes. The Nebulae software will also change RAID postures and N×M processor redundancy schemes to adapt to changing mission needs during critical events.

## D.1 Reliability, Availability, Fault Tolerance

**Reliability** measures the ability of a computing system to operate over a specific time interval without failure. For example, a system reliability of 0.9 for a 5-year mission means there is a 90% chance that this system is operational after 5 years. But say this system is impacted with non-destructive upsets due to radiation effects, and each upset requires a reboot to reload and repair the impacted memory. The system could be said to be reliable because these upsets are non-destructive and thus repairable, but the system may not be available all the time. **Availability** is a measure of a system to operate properly when needed. **Fault tolerance**, typically achieved through redundancy, is a way to improve availability by masking faults. Converse to the previous example, in a triplex processing system when a fault occurs, a majority vote of the 3 processors masks the faulty processor; therefore the overall system is available and processing science data even with one temporarily unreliable

processor. Redundancy also improves system reliability, as a single processing element can fail completely and the system will remain available and operational until the next fault.

## D.2 Radiation Considerations

Electronics in a deep space environment are exposed to energetic particles (i.e., protons, heavy ions) that can have non-destructive (i.e., bit flips, functional errors) or destructive (i.e., latch up, gate rupture) effects. There may also be device degradation over time due to cumulative radiation effects (total ionizing dose [TID], displacement damage, and enhanced-low- dose-rate-sensitivity [ELDRS]). Radiation-hardening by design (RHBD) devices can mitigate these effects at the device level, but at a significant cost. The cost is not only in dollars, as these devices are not inexpensive, but also in terms of efficiency (Watts/operation), which drives up power and thermal requirements for the mission. The lower efficiency compared to non rad-hard commercial electronics is primarily due to redundant circuitry added to address device upsets and the use or larger feature sizes on the die to mitigate the impact of each hit by an energized particle. Rad-hard devices lag commercial processing technologies by 10 years or more (see the CPU chart above) which is several generations of processing evolution and often more than one order of magnitude of processing capability. This presents a "Catch-22" or a dilemma: sacrifice performance to ensure radiation hardening or sacrifice radiation hardening for higher performance. To date, the solution has always been to sacrifice performance. The latter case is currently the area of scrutiny in the context of science application requirements.

## D.3 Radiation Revisited:Screening Commercial Parts

Rad-hard processors may have low operational efficiency when compared to commercial-off-the-shelf (COTS) electronics; and this in turn drives SWaP. COTS processors and circuits can be susceptible to both hard (permanent or non-recoverable) and soft (recoverable) faults. Therefore, to use COTS parts off-planet, the parts must be characterized through radiation testing. In most cases it is likely that failure modes will be uncovered (latch-up events, total dose degradation, functional upsets, bit errors, etc.). There will also likely be a fair amount of variability between commercial devices based on semiconductor processing technology, device layout, and even wafer lot. It effectively comes down to doing the due diligence to qualify and screen COTS devices for your application, which may require testing multiple devices to find one to meet performance requirements. To the extent possible, the failure modes uncovered during device testing must have mitigation strategies to achieve the required performance. Since the devices were not designed for the space environment, there is likely residual risk with this approach, but given the performance and SWaP benefits, the risk may be acceptable for some missions. There are best practices for selecting COTS manufacturers that should also be used to ensure reliable devices.

## D.4 Mitigation Strategies: Some Examples

Using COTS devices that have been characterized through radiation screening must meet basic requirements (i.e., lifetime requirements and the ability to manage destructive fault modes). Assuming this is the case, additional mitigation may be required. Various techniques are available for designers to mitigate radiation effects. The mitigation strategies may vary for different applications, but they can be thought of as adding robustness to different levels in the processing architecture (see Figure D.1).

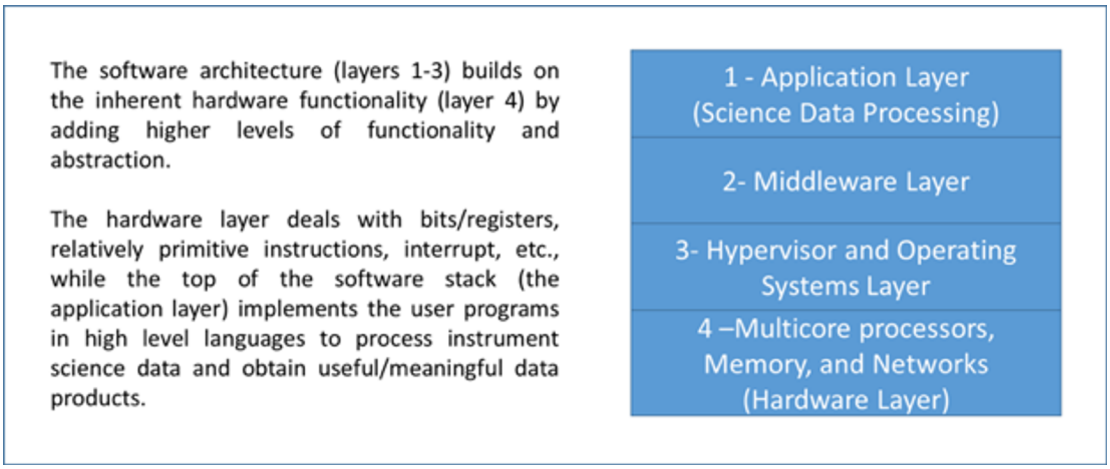



Figure D.1: Hardware and software architecture.

At each level in the processing hierarchy, different mitigation techniques may apply. Figure D.2 identifies potential mitigation strategies that are appropriate for different levels in the hierarchy. This list is not exhaustive and in some cases not all techniques will be utilized. In addition, the details of these mitigations are beyond the scope of this paper.

## D.5 Mitigation by Shielding

Radiation effects are also mitigated by bulk shielding, molded shielding, incidental shielding by using interior locations in the vehicle assembly, and soon will use additively manufactured shielding tailored precisely for each vehicle. These topics were taken as a given for Nebulae and are researched elsewhere.

No matter what mitigation techniques are employed, however, truly strange things happen in high-radiation environments, and the system designers must assume the worst will happen.

## D.6 Storage Resilience

Another good example of scalable resilience can be seen in the form of terrestrial high-reliability file systems such as ZFS `https://docs.oracle.com/cd/E19253-01/819-5461/zfsover-2/`. ZFS has the ability to adjust the number of single bit and device level failures that a

system is tolerant of by adjusting the count and topology of the non-volatile storage elements that make up the overall file store. These same techniques can be applied to solid-state data recorders to develop high reliability storage that is built on top of unreliable but high-density COTS flash devices. Because the topology of the storage can be adjusted in software, different levels of reliability and redundancy can be created on an as-needed basis.

## D.7 Conclusion

The science community has a need for high-performance computing systems with reasonable size, weight, and power to be implementable on deep space missions. To satisfy these constraints simultaneously, a dynamic Resiliency as a Service can be employed that allows use of advanced COTS devices in addition to rad-hard devices, which can realign processors from single-string to parallel strings depending on mission need, adjust the fault tolerance of storage systems, and dynamically circumvent radiation events.

## D.8 Processor Redundancy and Fault Tolerance

An overview of various processing approaches is shown in Figures D.3 and D.4.

These figures show the trade space of typical processing approaches examining fault tolerance, availability, and power. Reliability is not included as it is assumed that part selection, radiation performance, and mission life will be the primary drivers for reliability. For now it is assumed that all systems could meet a minimum reliability requirement—this will be revisited later.

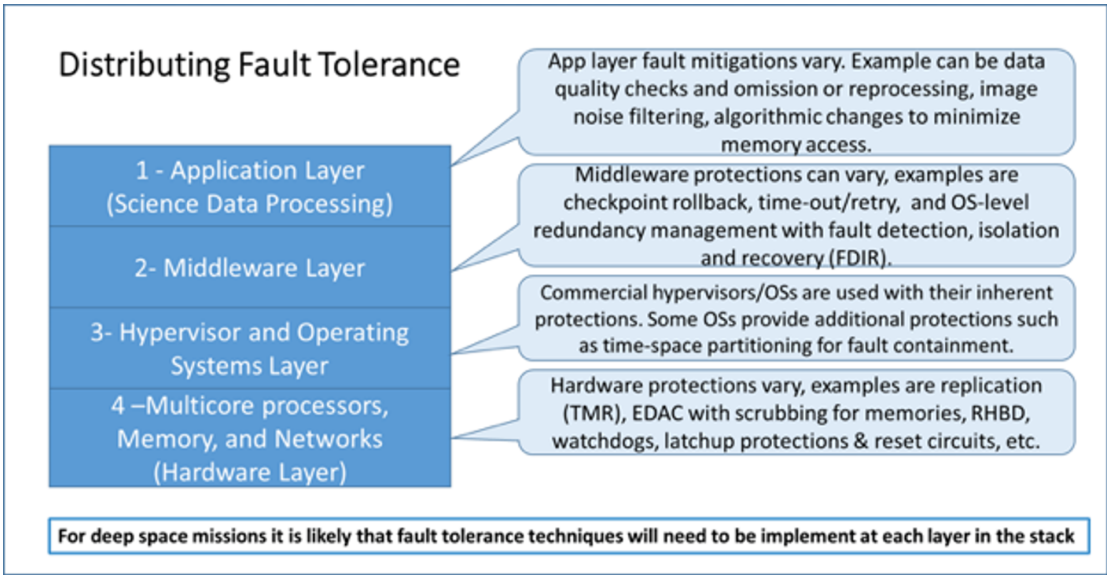



Figure D.2: Potential mitigation strategies at different processing hierarchy levels.

| Processing Approach | Fault Tolerance (see notes) | Availability | Power (level) |
|---|---|---|---|
| Reboot on Fault | Note[1] | Time to reboot and redo computation after a fault | Low (baseline) |
| Checkpoint/Restart | Note[2] | Time to reboot and continue computation after a fault | Low-Medium (x1.5) |
| Primary/Backup | Note[3] | Operates through a fault | Medium-High (x2) |
| Voter or Self-checking Pairs | Note[4] | Operates through a fault | High (x3-x4) |

Figure D.3: Trade space of typical processing approaches.

At the highest level, Figure D.3 shows that the power requirement increases with increases in fault tolerance and availability. Along with the power increase, weight and size would also increase. Seeing this trend leads to the question, "For science data processing, is it acceptable to have lower fault tolerance and periodic loss of availability to reduce SWaP?"

| Processing Approach | Fault Tolerance Notes |
|---|---|
| Reboot on Fault - Note[1] | Faults that stop processing must be detected and the system is restarted from the beginning, since intermediate computations have not been stored. With this approach it is possible that certain faults go undetected (eg. computational values in a calculation is changed). |
| Checkpoint/Restart - Note[2] | Faults that stop processing must be detected and the system is restarted from the last saved checkpoint. With this approach it is possible that certain faults go undetected (eg. computational values in a calculation is changed). There could be significant overhead with checkpointing based on frequency but less overhead than multiple processing elements. |
| Primary/Backup - Note[3] | A detected fault in the Primary processing element will cause the Backup element to take over as primary. The failed processor must be recovered to prevent additional faults from causing a loss of capability. With this approach it is possible that certain faults go undetected (eg. computational values in a calculation is changed). |
| Triplex Voter or Dual Self-checking Pairs - Note[4] | A detected fault in a processing element will be masked by other processing elements. The failed processor must be recovered to prevent additional faults from causing a loss of capability. Virtually all types of faults are detected by this processing approach. |

Figure D.4: Fault tolerance notes on typical processing approaches.

Science data processing involves the examination of sensor data to look for physical phenomena (i.e., water/ice, geysers, dust devils, etc.) and report this information back for further analysis. When availability is reduced, there is a greater possibility of missing a scientifically significant event. But in the course of a long duration mission, losses in availability will be small when compared to the mission duration and therefore acceptable. Gaining significant overall sensor processing—as much as an order of magnitude more ground surveyed—in exchange for gaps in the processed data, on the order of 1% or less, is a similar trade study that drives processor selection and fault tolerance architecture.

Through this analysis and based on inputs from the science community, the simplest single-string architecture was selected knowing that rebooting on faults may be required to restore functionality. This statement has the implicit assumption that all faults are recoverable to maintain reliability—to ensure that this is the case, the topic of radiation needs to revisited.

There are nuances to fault tolerance, different classes of faults requiring varying levels of redundancy, for example. But for the purposes of this discussion, these can be largely overlooked. The question to answer is, "What are the minimal reliability, availability, and fault tolerance requirements for a science data processing system?"

For science data processing, detecting/dealing with faults will vary by application. The range of impacts will be broad; from a major system fault to a single bad pixel to missing detection of a change. Fault recovery approaches in software will also vary across the full spectrum: Do nothing (e.g., live with a bad pixel); redo a unit of computation; restart a unit of computation from a checkpoint; or use some state recovery technique to move a software process if necessary and resume where the application left off.