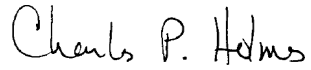
Ad-Hoc Task Force on Big Data
of the
NASA Advisory Council Science Committee

Meeting Minutes

September 28-30, 2016
NASA Ames Research Center



Charles P. Holmes, Chair

Erin C. Smith, Executive Secretary

Report prepared by Joan M. Zimmermann
Ingenicomm, Inc.

Table of Contents

September 28, 2016

Introduction

Dr. Erin Smith, Executive Secretary of the NASA Advisory Council (NAC) Ad-Hoc Task Force on Big Data (BDTF), called the third meeting of the Task Force to order and provided administrative details for the meeting's duration. She updated the committee on the progress of the Terms of Reference (TOR), and informed the BDTF that it will be extended to January 2018, to match the expiration date of the initial membership.

She introduced Dr. Charles Holmes, Chair of the BDTF. Introductions were made around the table. Dr. Holmes provided details of his report, including the fact that NASA has announced a new Associate Administrator for the Science Mission Directorate, Dr. Thomas Zurbuchen, a heliospheric scientist from the University of Michigan. Dr. Holmes was looking forward to his tenure and was in the process of informing him of the BDTF's latest activities. Two recent candidates for expanding the membership of the BDTF fell through, and given the timing, there will likely be no further pursuit of candidates. Dr. Holmes described his report to the Science Committee (SC) in June 2016 as having been well received. The four findings, including some informal comments from the SC, were briefly reviewed. The SC passed BDTF's finding about the Astrophysics Data System (ADS) on to the NAC, members of the SC agreed that ADS is an underappreciated resource. The SC requested feedback about the evolution of archival data formatting and also provided some good commentary on the use of the Flash I/O Network Appliance (FIONA) and its readiness for widespread adoption. Dr. Reta Beebe felt that a different task force might be needed to assess the utility of FIONA, as it seemed too broad an issue for the BDTF. Dr. Holmes felt that this suggestion should be aired going forward, but generally thought that the BDTF needs to be more careful about what it takes forward and instead work on larger, more strategic questions. He felt also that the BDTF's previous findings were too verbose and should be more concise in the future. He listed meeting goals: to hear about archiving of astrophysics data at Caltech and large-scale computing; to hear about Ames projects focused on Big Data technology; to develop momentum on task force studies going forward; to tour the Ames Research Center and Big Data centers in Palo Alto; and to generate findings and recommendations (F&R) for the SC this Fall.

Dr. Holmes briefly addressed the ongoing NAC SC reorganization, in which subcommittees were being elevated to report directly to the NAC. Dr. Smith commented that functionally, the reorganization shouldn't change BDTF proceedings. Dr. Smith further announced she would be moving to Goddard Space Flight Center to serve as the Deputy Observatory Scientist for the James Webb Space Telescope (JWST) mission, and would be succeeded as Executive Secretary by Mr. Gerald Smith.

Welcome to Ames

Dr. Tom Edwards, Deputy Center Director, provided an overview of the Ames Research Center (ARC)/Moffett Field, a former naval air station. Over the last 15 years, NASA has been repurposing space at Moffett Field, having leased out the airfield and hangars to Google Planetary Ventures, LLC. ARC is comprised of 2000 acres, and employs 1150 civil servants and an equivalent number of contractors. It operates on a $750M budget, in addition to a fair amount of reimbursable work. ARC houses the largest wind tunnel and largest motion simulator in the world, an arc-jet laboratory, and a supercomputing center. The center has a diversified research portfolio and supports all NASA mission directorates, particularly aeronautics, air traffic control systems, aerosciences, entry and descent systems, and thermal

protection systems (TPS). ARC also supports modeling and simulation of environments, space and Earth sciences, the Airborne Science program, autonomy and robotics (intelligent/adaptive systems), advanced computing and IT systems, astrobiology and life sciences, and the Cost-Effective Space Missions program (nanosats, cubesats and smallsats). Kepler and the Lunar Atmosphere and Dust Environment Experiment (LADEE) are examples of these latter reduced-cost missions. ARC served as a pilot center for leasing property to tenants interested in collaborating with NASA, while assisting in technology transfer and related tasks. Today ARC regularly collaborates with academia, small business, and commercial entities. ARC is now 100% occupied by tenants, including Moon Express, Bloom Energy, and Made in Space, representing many success stories in the space industry.

Asked about the iconic Hanger One at Moffett Field, Dr. Edwards explained that it would be re-skinned after environmental remediation has been conducted. Planetary Ventures signed up to do the remediation and is currently looking at various architectures. The other two large hangars (Two and Three) on the property are also being refurbished by Planetary Ventures. They are wooden and require fire suppression systems. Dr. Holmes asked if the government possessed any liens on intellectual property developed under agreements. Dr. Edwards explained that tenants have a straightforward lease, but in many cases there is a Space Act agreement that directs shared intellectual property (IP) in the non-reimbursable cases. Ames is known for crafting such partnerships, as in a current agreement on quantum computing. The World Wind program, e.g., was developed at Ames. He also cited an instance of ARC's utility to the Kepler mission, which requires much computing power to determine whether transits represent planets. At one point, incoming data was outpacing the mission's ability to reduce it. As a solution, the mission ported the data over to ARC's Pleiades supercomputer and solved the problem. Such synergies are everywhere within the agency; keeping pace with data assimilation is imperative.

<u>NASA Big Data Challenges: Ames Perspective</u>

Dr. Piyush Mehrota, Chief of Supercomputing, gave a briefing on the Ames approach to Big Data. The center's core capabilities, several of which touch on Big Data, include Earth Science, Aerosciences, Astrobiology and Life Sciences Core, and Advanced Computing. The National Strategic Computing Initiative, an Executive Order, is helping Ames to increase the coherence between the technology base and its users.

Ames houses the High-End Computing Capability (HECC) project, NASA's premier supercomputing center, which serves all mission directorates of NASA and all subdivisions of SMD. HECC includes over 500 projects and 1500 users. Its biggest system is Pleiades, a distributed memory cluster that operates at 7.25 petaflops at peak usage. Pleaides ranks number 15 in the world and number 7 in the US, and its High Performance Conjugate Gradient (HPCG) benchmark ranks at number 9. Although Pleiades is #15 in the world, NASA is always bringing in new hardware for testing, and can use four generations of systems. Ames functions as the "smart buyer" for the agency, and provides a balanced environment of computing for both science and engineering. The system is dumping 2 petabytes of data per month. Asked how the results of specialized hardware tests were broadcast, Dr. Mehrota said that this was done through papers published in supercomputing journals, as well as through webinars on how to use environments, and best practices. These results are broadcast widely, well beyond the 1500-user community.

HECC allocations are controlled by mission directorates through integrated spiral support for modeling, simulation and analysis (MS&A). The center provides a Level 2 help desk that delivers the most productive, integrated supercomputing environment in the world, and

includes a group that does visualization (NASA Hyperwall). The focus is on large data sets. Big data challenges arise from the enormous number of satellites, telescopes, and missions, totaling about 50PB of data per year. This includes observational, model and experimental data. The center is getting more and more simulation runs, as well as experimental data from complex wind tunnel experiments, e.g. These experiments use high-resolution video cameras to observe changes on surfaces covered with pressure-sensitive paints, using continual feedback to reposition the cameras as they observe the color changes.

The challenge of Big Data is not just analytics; because many people don't know where the data is. Capabilities such as data discovery and searching metadata, and searching across data types, are becoming increasingly important. Data management, transferring large data sets from the archives to computational resources, managing complex work flow— all of these tasks need software to help manage them automatically. Dr. Clayton Tino asked if a standardized tool set was being used internally. Dr. Mehrota reported that this was not yet the case. At the infrastructure level, I/O is becoming more important, and Ames is working on trying to get this in place. Large memory spaces for in-core analysis, support for heterogeneous resources, and data dissemination (how to engage scientists outside the projects) are all foci of the HECC project. The key point is that it is the whole pipeline. Dr. Mehrota noted that the term "Big Data" was actually coined at Ames, at a time when its biggest data set was 7.5GB.

The merging of analytics and high-performance computing (HPC) are leading to a larger vision of how to make Ames a good center for data analytics. This will require a high-speed network, and a determination of how to ideally merge analysis platforms, data resources, and large-scale computing. Big Data-related projects at NASA include mining network flows for malicious events, tree cover classification for the continental US, data tagging for security and data discovery, ontology-based data search environment for observational data, looking at SSDs for I/O optimization.

In summary, NASA has an abundance of big data, and the challenges are wide. Ames is the ideal location for merging data analytics and HPC. Dr. Holmes commented that over the next year, the BDTF would be working on several topics, and might want to drill down deeper and come up with a NASA/Ames story. Dr. Mehrota was happy to answer any questions on that subject and cited the recently retired CTO, Deborah Diaz, as a good resource. The NASA Big Data Working Group is co-chaired by Dr. Tsengdar Lee and Dr. Diaz. Dr. Holmes asked Dr. Lee to give an impromptu introduction to the Working Group at the next meeting of BDTF. Dr. Mehrota pointed out that NASA should focus on structured data, as opposed to what the commercial world is concentrating on. Dr. Tino noted that real-time streaming and distribution of data represents one area that is able to leverage more commercial solutions. He felt that wind-tunnel experiments with pressure-sensitive paint, for example, were good test cases for using commercial models that use commodity compute, which is cheaper than using a supercomputing center. Dr. Mehrota noted that Pleiades is a private, highly secure Cloud, by contrast to commercial Clouds; Pleiades is effectively a Cloud that doesn't use commercial hardware Dr. Mehrota felt it didn't make sense for Ames to use the commercial Cloud for HPC. Dr. Lee commented that NASA had conducted some experiments on trying to offer infrastructure as a service, noting that many people want to bring their own platforms to HPC. The two types of workloads are not quite compatible. Dr. Tino said he was just bringing up a concern that HPC is not considering super-cheap hardware, while understanding that platform-as-a-service does not necessarily make sense for NASA.

<u>NASA Ames Data Sciences Group</u>

Dr. Nikunj Oza, Leader of the Ames Data Sciences Group, presented a briefing on his group, which is oriented toward data mining research and development (R&D) with applications to NASA problems. The group comprised of 11 members and is funded by Aeronautics, Earth science, and some Space Exploration and non-NASA funds. In Aeronautics, the focus is mostly anomaly detection, precursor identification, and text-mining for identifying problem topics. In Earth Sciences, one problem is to fill in missing measurements. The group also studies graph-mining algorithms, which originally developed to monitor e-mail traffic patterns to identify potential threats, and now applies them to understanding the climate. For the Kepler mission, the group worked on refining algorithms for planet-finding, and also on algorithms for monitoring astronaut health. Of the "Four V's" of data challenges, Volume is the least challenging, but Velocity is a big one that needs to be refined. Veracity is also a challenge: in air traffic control, examples are duplicate tracks, data drop-outs, and tracks ending in midair. Variety is another challenge: numerical data (binary vs. continuous); weather (forecast vs. actual); and radar/airport data. Aeronautics data mining problems include anomaly discovery over a large set of variables; identifying particular variables of interest, e.g., fuel burn; precursor identification (preventing go-arounds); and topic extraction such as crew fatigue, late hours, etc. Generally with Aeronautics, the method is to use exceedances, which doesn't necessarily find anything new. Dr. Oza felt that data-driven methods let the statistics of the data speak for themselves. Although this approach can result in false positives, it can also identify new problems, such as strange shapes of trajectories that can lead to safety issues. The Data Sciences Group is also working on providing domain expert feedback to anomaly detection algorithms to reduce their false positive rates.

In an Earth Science example, researchers seek to understand sensitivities to Amazon droughts, and vegetation anomalies using genetic algorithms and symbolic regression trained on massive Earth science datasets with different resolutions and projections. Using seasonal variables and smoothing, there has been some progress in reducing errors. Ongoing and future work will involve deriving nonlinear models on Amazon tiles and improve scalability so that the algorithms can work on a global scale. VESsel GENeration (VESGEN) is another ongoing project, which models vasculature of the eye. VESGEN is trying to speed up/automate a process that usually requires 10-25 hours for producing a single retinal image, by using more pre- and post-processing techniques, and machine learning, to solve the problem.

Dr. Oza closed by mentioning DASHlink, a collaborative website designed to promote interaction, and that he was also a member of the Big Data Working Group.

<u>NASA Earth Exchange (NEX)</u>

Dr. Rama Nemani presented an overview on a six-year-old Ames project, the NASA Earth Exchange (NEX), which deals with massive data from multiple missions using scalable and diverse computing architectures. NASA has an established business model in the Earth Sciences (ES) that is driven by specific science questions; i.e. what sensors, what measurements to take. Now, ES is trying to morph into integrating data from different sensors and answering science questions in different ways; each discipline addresses its questions in its own way, and is asking more complicated questions than in the past. NEX arose from a need for an ES collaboration. Earth Science is a community effort involving 179 institutions in the US, and hundreds of investigators. Redundant storage and processing facilities require larger budgets to move data sets that are getting larger each year. Sharing knowledge is becoming more difficult.

NEX is a virtual collaboration that provides a complete work environment, with roughly 2 PB of data in a centralized repository. It creates a portal and an attractive platform for people to

process large-scale data. NEX resources include a portal, a sandbox for testing codes, and finally the HPC environment. Codes and models include GEOS-5, WRF, BEAMS, TOPS, etc. Projects include high-resolution climate projections for climate impact studies; high-resolution monthly data for monitoring forests, crops, and water resources; and mapping fallowed areas in the California drought. Moving forward, NEX will be supporting machine learning and data mining, and is moving toward more data-driven approaches; this has resulted in refinement of such things as Moderate Resolution Imaging Spectroradiometer (MODIS) fire data. OpenNEX is a private-public partnership that uses a commercial public Cloud to provide labs, lectures, and workshops—users pay only for computing time through an Amazon Cloud. Dr. Lee asked if anyone has been paying Amazon without NASA in the loop? Dr. Nemani said, absolutely yes. In summary, the main idea of NEX is to lower the barrier for testing ideas, share knowledge, and to provide transparency for climate studies (as it is a controversial topic), to yield reproducible, verifiable results.

Dr. Petr Votava presented NEX technology goals: supporting current and future diverse large-scale Earth Science through science data management; workflow/process management (for monitoring progress and provenance of data); outreach and engagement; and knowledge management. NASA would like to bring other communities in to share and collaborate in data, coding, training, etc. Challenges include system access; NEX approached this challenge by setting up a sandbox outside the secure framework to encourage broader engagement with data. NEX architecture can accommodate non-NASA users, while they are acquiring NASA approval, through the sandbox; when they are finally approved they are ready to go. NEX provides automated software to make it easy for users to move code, and is addressing processing challenges through the use of agile processing pipelines, automated provenance and reproducibility; improving I/O bottlenecks and inefficiencies; and experimenting to anticipate future technology trends.

NEX science analytics challenges include the effort of balancing cost and efficiency. Most data are not "analysis-ready." NEX is experimenting with science analytics architecture and working with science databases. Classes of NEX Big Data projects include fully distributed data processing with no interprocess dependencies; machine learning and data mining with some interprocess data dependencies; and distributed independent data processing. Analysis and science applications include such things as mapping crop water requirements, provenance and knowledge graphs, and climate and ecosystem modeling. In progress is the BEX Big Data Project, an interim project that deals with long-term Landsat data processing, whose is goal to create consistent data for observing Earth changes, using a WELD-GIBS processing pipeline. The entire NASA data holdings for Landsat constitutes 16PB. The pipeline will push through 25PB for 18 years of Landsat data. Dr. Tino addressed the issue of repetitive processing of small data sets, and asked if NEX were aligned with other efforts or services that can solve the problem. Dr. Votava said that NEX is currently addressing how to get to the data (at a lower level), as it's difficult to put the data all back together. The goal is to be able to move data somewhere else to reproduce results. Dr. Tino commented that he has frequently seen NASA scientists struggle with entrenched habits, and recommended some evolved thinking on what "accessing data" means.

Dr. Sangram Ganguly presented recent NEX efforts on solving tree cover classification problems, which is in the testing and prediction phase. Delineation is a hard problem because tree cover is represented by random patterns, and the quality of data is affected by many variables. NEX has had to derive data from 330,000 National Agricultural Imagery Program (NAIP) scenes, and it took many months to even transport the data. As of 27 September, NEX

had finished processing ten US states. NAIP processing architecture was described as typical Hadoop-type processing. Dr. Ganguly displayed some experimental results on the California Tree Cover Mosaic, which will eventually expanded to the remainder of the continental US, and a brief slide on DeepSAT, which is a learning framework for satellite imagery. Dr. Ganguly encouraged people to use SATNet, the largest satellite imagery dataset and model zoo, to help efforts along.

Dr. Holmes encouraged a continuing dialogue with the NEX group. Dr. Mehrota offered himself as the first-level contact. Asked if Ames was engaged with the Goddard Institute for Space Science (GISS), Dr. Mehrota indicated that most interaction was with GSFC. A meeting participant interjected that there is in fact an Ames person engaged with GISS. Dr. Mehrota noted that the biggest issue has been with security policies when engaging non-NASA participants. Federal policies are very inflexible. Dr. Holmes lamented that he'd been hearing this complaint for 20 years. Dr. Lee commented that government tends to lock down everything; NASA needs to assess risk differently for science vs. space systems. Dr. Raymond Walker commented that he had also been fighting the problem with the Planetary Data System (PDS), despite the fact that it is an open data set. Dr. Holmes flagged security as a topic for discussion and possible findings. He felt that SMD should champion the issue. Dr. Smith commented that cybersecurity is more of an Office of the Chief Information Officer (OCIO) problem. Dr. Tino recommended a finding that highlighted the current security posture of the Agency and how it hinders collaboration. Mr. Smith also indicated that the OCIO is actively dealing with the matter.

Lunch talk

Dr. Pam Marcum presented a lunch talk on science results from the Stratospheric Observatory for Infrared Astronomy (SOFIA) mission.

Public comment

No comments were noted.

Supercomputing in the Age of Discovering Superearths, Earths and Exoplanetary Systems

Dr. Jon Jenkins presented an overview of exoplanet discovery, noting that there are now 3338 confirmed exoplanets as of September 2016. The first exoplanet was discovered in 1989, and initially it was not clear whether or not it was a brown dwarf. 51 Pegasus was found in 1995. Methods for discovery include microlensing, imaging, non-Kepler transits, and Kepler transit monitoring. In 2000, an "unequivocal" exoplanet was found based on its radial velocity. Over time, sensitivity using radial velocity got better and better. Today, researchers are finding Earth-sized planets via radio velocity methods. Proxima B was recently discovered in this manner. Kepler has found 2000 planets out of the 3000-plus known, while others are found via ground-based (GB) observation. Most planets discovered by GB transit are Jupiter-sized. Kepler allowed us to increase field of view (FOV) and sensitivity. These very large planets are curious objects, and their composition is not yet known. In terms of enabling Kepler, back-illuminated CCDs with high quantum efficiency, sophisticated algorithms for identifying weak signatures, and significant computational infrastructure have been key.

Kepler works by staring at a large FOV over time, watching for variation in star brightness. Duration of the transit helps constrain the orbital period of the candidate exoplanet. A Jupiter-sized planet transiting a sun-sized star typically causes a 1% variation, while an Earth-sized planet results in a 0.01% variation. Processing data from Kepler involves calibration of image data; measurement of brightness (photometric analysis); removal of image artifacts (pre-search

data conditioning); input to the Transiting Planet Search to flag Threshold Crossing Events, which are then validated and reviewed by a review team. NASA is applying machine learning to the candidate evaluation process to supplement the efforts of the Threshold Crossing Event Review Team (TCERT).

Kepler has started porting software to the Pleiades, with results broadcast to the Hyperwall, which can in turn be used for parametric studies. One result of such an analysis was the confirmation of the discovery of the Earthlike Kepler-452b that is thought to have Earth-like features. Transit-like signals can also be produced by background eclipsing binaries, triple-star systems with an eclipsing binary (EB)/planet, and background/foreground planet systems. The BLENDER program can assess statistical confidence for the planetary nature of a candidate; it is computationally intensive and requires a supercomputer. The BLENDER program was applied to Kepler-452b and helped cap the peer review of the discovery.

Other supercomputing efforts are being put to bear on searches for exomoons in 400 light curves from Kepler. Each search consumes 50k CPU hours. To date, no exomoons have been conclusively discovered. In the future, the Transiting Exoplanet Survey Satellite (TESS), which will look at about a 200-light-year radius, will identify the 50 best targets for follow up by JWST. TESS is scheduled to launch in December 2017, pending amelioration of problems with the Falcon 9 launch vehicle. NASA plans to analyze TESS data with supercomputing resources from the start. Supercomputing is deemed as absolutely essential for the TESS mission. Dr. Beebe asked if the entire Kepler CD had been read out yet. Dr. Jenkins reported that not all of the CD had been read; since the mission was repurposed in 2014, there have some limitations associated with reduced bandwidth and limited onboard storage. Similar pipelines are used for astroseismology, and have been every bit as successful as the processing used for Kepler.

Dr. Eric Feigelson commented that ES typically processes petabytes, while Kepler represents basically 1TB for light curves. He asked why supercomputers needed for these relatively small data sets. Was it for sophisticated statistical modeling? Dr. Jenkins replied that yes, the computational intensity of the algorithms require HPC. Dr. Lee had the impression that the Kepler pipeline did not take advantage of multiple processors. Dr. Tino likened the process to batching out MATLAB jobs, where the processing is a function of the batch scheduler. Dr. Mehrota felt that something like this would fit very well in the Cloud. Dr. Jenkins noted that each of the individual pipelines are different, so load balancing becomes an issue. Some stars process quickly, and some take a long time. It's a system software problem. Dr. Feigelson asked if there were an interest in the TESS team to move on to the Cloud, perhaps as a test case. Dr. Jenkins deemed this an interesting question; at the preliminary design review (PDR), TESS adopted the design from Kepler; the mission looked at the Cloud in 2015 and at the time did not feel it appropriate. To change the design at this point would be too costly. Dr. Tino asked if TESS data processing tasks were to be scheduled as one big job. If this is not the case, he felt it would be worth exploring the concept of scheduling many small, simple jobs to the Cloud. Dr. Mehrota thought that this assessment was essentially correct, but that implementation was crucial to success; that's why NASA needs to look at ways to evolve missions along with the computer technologists. Dr. Neal Hurlburt commented that small missions typically don't do the data systems until the instruments are built, so as to allow for adaptation of the most current technology. Dr. Lee noted that if this type of job keeps coming to the supercomputing center, there may be a need to maintain two platforms: one for highly coupled applications and one for simple data pipelines. Dr. Tino observed that the most successful projects seem to combine the mission scientists with the computational scientists. Dr. Holmes agreed that this was a behavior modification issue. Dr. Tino felt it to be a strategic issue as per the interest of the Science

Committee. Dr. Hurlburt added that the subject might also be of interest to the Workforce Task Force.

Member Reports/Discussion
Members presented their individual reports:

Dr. Feigelson reported that in astronomy, there tends to be significant support from philanthropic groups, or commercial concerns such as the Discovery channel. He noted that one example was the Sloan-Moore Foundation, which is giving a total of $35M over 5 years to support HPC for astronomy at the University of Washington-Seattle, New York University and University of California-Berkeley. The foundation's goal is to systemically change the way researchers work. In addition, the Simons Foundation, founded by an NAS mathematician, is making similar investments to advance the role of mathematics in astrophysics. Astronomer David Spergel is the director of the foundation, home to a total of 60 PhD/professional level members with expertise in high-performance computing in astronomy, in the areas of Big Theory and Big Data. NASA should be aware of these efforts.

Dr. Feigelson further reported on two astroinformatics meetings scheduled for October 2016: the 26th Astronomical Data Analysis and Software System (ADASS XXVI) conference. ADASS has the strong presence of the European Space Agency (ESA) but not NASA. The other meeting is the International Astronomical Union Symposium (AstroInfo 2016), which is interested in pushing the envelope. He noted that there was also no NASA participation in this symposium, and very few NASA-funded individuals.

Dr. Beebe reported that in 2006, motivated by ESA, an International Planetary Data System (IPDS) was formed to undertake a major effort to look at specific problems to be integrated across international agencies. The IPDS meets face to face once per year, and holds monthly telecons. The group has a Steering Committee and technology advisory groups, which produces white paper reports. At its last meeting in July 2016 in Madrid, IPDS adopted the use of NASA's PDS4 data standard. Most international space agencies will now use the standard, including JAXA and the United Arab Emirates (UAE). Dr. Beebe added that there is now a memorandum of understanding (MOU) with North Korea to support a lunar mission. The PDS4 standard now includes a registry and search system, on which one can register old data. For users want to search on physical parameters, the purpose of PDS4 is to help facilitate that search. She added that the IPDS also enjoyed good collaboration with the Ames ultraviolet group.

Dr. Hurlburt noted that he would be going to the ADASS meeting mentioned by Dr. Feigelson. He reported the launch of a new search tool for internal missions (e.g., IRAS, Hinode), and noted that once the tool launched, the use of old data was seen to increase.

Dr. Tino reported having spent recent time in discussions with his internal data group, exploring such things such as Apache Storm, and the feasibility of using bundled analytics tools in the Cloud.

Dr. Walker reported having contacted Larry's Smarr network CIO in order to make use of UCLA's 100Gb/s pipeline, which is so new it hasn't yet been made widely available. Dr. Walker has volunteered to be a first user. A UCLA Chemistry faculty member will work with him on a 40Gb/s connection. He is also testing bringing Pleiades simulation run data over the 100Gb line. In recent testing, given that the pipeline is brand-new equipment, Dr. Walker noted that the hardware malfunctioned initially. The second attempt worked very well. The previous day, he

noted that the pipeline had been tested all the way from Pleiades, where it ran into software problems. It was re-tested and Dr. Walker thought it had succeeded at getting the speed to about 1Gb/s. Dr. Walker also computes at Kyoto University and now has an account at the Pacific Rim Consortium, which he offered to report on at the next BDTF meeting.

Dr. Holmes reported the National Science Foundation (NSF) Regional Innovative Hubs project, for which the Spokes competition has concluded. Results are available on the NSF website. There were 10 $1M awards, and another 10 $100K awards, spread evenly across the four geographic regions, and ranging over a great span of topics. He requested that BDTF members go back and talk to their contacts to determine where they're heading, and maybe host some NASA science researchers, after which the TF would write up results at its February 2017 meeting.

September 29, 2016

Enabling NASA's Use of Cloud Software-as-a-Service (SaaS)

Dr. Raymond O'Brien presented a briefing on how NASA is facilitating access to the universe of Cloud products for NASA employees. He believed the model would change the way NASA computes, and noted that industry analysts agree that Cloud use has been a disruptive force in the enterprise IT industry. Another analyst believes Google's Cloud capability will reach a trillion dollar value in due time. Cloud computing is changing the service delivery model from "pallets and crates" to seconds of transfer time. When the smoke clears, there will be just a handful of service providers. Software-as-a-Service (SaaS) has few barriers to entry; as it can deliver product with very little investment. NASA needs to access it in a secure responsible way. To accomplish this, NASA has formed Enterprise Managed Cloud Computing (EMCC) to put the pieces of the puzzle together to facilitate the workforce's access to Cloud service, given that the demand for SaaS in the near future is expected to represent five times the demand of traditional installed services.

The case for an enterprise SaaS approach does present some challenges; different projects may interpret requirements differently, and there can be unknown security postures, with associated risks. Long-term goals of the EMCC include the widespread adoption of Cloud computing by programs and projects. NASA's intent is to use commercial SaaS to address requirements for programs and projects when it is the best approach.

NASA runs many applications; Cloud commercial apps for NASA are expected to expand greatly in both technical and administrative areas (payroll, etc.). EMCC therefore needs to facilitate permissions for some apps, such as body-scan based designs to improve the fit of astronaut suits. The key to strategy is the service delivery platform. EMCC's goal is to reduce the use of unsanctioned services; improve Agency risk posture; bring IT services under enterprise management; and provide a unified-services delivery approach that will enable innovation to address mission requirements.

A platform is needed because growth will be explosive. There are at present 20,000 SaaS providers, with some products that are not suitable for NASA information. Unsuitable products need to be identified. NASA wants to leverage bundles and discounts as much as possible. Key elements of the SaaS strategy include deployment of an Agency-wide SaaS marketplace, and development of a network of service providers, giving precedence to early SaaS apps that can be onboarded to a "Minimal Risk Portfolio" (MRP), which will not require advanced security vetting.

Using a tiered-service architecture, EMCC uses managed Cloud environments (MCEs) to help manage access across the Agency. An implementation strategy is under way; security processes are being put in place for NASA employees, while NASA develops tools to indicate the risk posture of the apps; the Agency is leveraging the current identity system to do this. Once the Cloud Access Security Broker (CASB) tool is deployed, "best of breed" of apps will be made available to NASA. EMCC will establish ownership of each of the applications, which will typically be groups that are already making use of a certain product.

Determining the business environment will take a little more time, as industry will help with its increasing number of tools. SaaS will have a life cycle: identify, authorize, onboard, monitor security, backup, and retirement. CASB brokers will sit between NASA and the universe of SaaS products, categorize the products, provide the risk postures, and provide snapshots of how much data is being used. EMCC is also watching potential developments in SaaS Aggregation Platforms, as NASA would prefer not to develop a platform in-house.

FedRAMP is the policy that guides access of federal employees to the Cloud. The key to enabling an effective security approach is through SaaS accreditation and authorization. There will be evolutionary changes in the EMCC operating concept from Generation 1.0 to 2.0, thus NASA is going to need a network of internal Cloud network providers, which will probably reside within OCIO.

Dr. Walker commented on the PDS infrastructure hosted at UCLA, which has gotten busy over the years, adding that there's an associated maintenance cost, using commercial software. He asked how one might make the decision that is most cost-effective for moving a NASA system to the Cloud. Dr. O'Brien recommended a technical cost estimate to uncover non-obvious costs.

<u>NASA World Wind</u>

Dr. Randolph Kim presented NASA's solution for seeing things in a global context, World Wind (WW), a 3D globe app that is based on open source software. World Wind is a software development kit that allows users to build their own 3D globes on a number of platforms, including Java, Android, and HTML 5. Dr. Kim displayed a sample globe on an Android device. The source code, available on GitHub, visualizes large volumes of data. Defense and government user communities are already using it widely; e.g. for command and control systems using World Wind as a background, or to see air traffic. The Department of Defense (DOD) has some hardened devices that employ WW as well, that will be deployed in 100,000 Army vehicles, to do such things as render radar tracks, and determine friendlies vs. unfriendlies. WW can be used to project video on the surface of the Earth, yielding 3D representations of unmanned aerial vehicles (UAVs). It is also being used to help visualize the FAA's next generation National Airspace System (displaying flight numbers, weather conditions, etc.). WW can be used with JSat to calculate trajectories of the catalogue of satellites, or to view satellite constellations such as the A Train. Another example is the Wildfire Management Tool, which incorporates parameters such as dynamic prediction, vegetation, slope, wind, humidity, and sun. The tool allows firefighters and emergency responders to visualize where fire fronts are heading. It is a very impressive tool, now available for mobile devices. The LIDAR Point Cloud, developed at the University of Kansas, loads raw lidar data and generates terrain. A Virtual Ocean has been developed by Columbia University, using WW. ESA uses WW as well; NASA and ESA are working on formulating a partnership to develop more applications. JAXA has used recently used the app for outreach for lunar displays.

Dr. Walker asked if people come to WW with a particular problem, or if one could just download the package. Dr. Kim said both situations have occurred, as the whole community is coming to realize they have common interests. Dr. Lee noted he had tried to use WW a couple of times and found it overwhelming. He asked if there were plans to package WW as an app with limited functionality. Dr. Kim cited the Android package as one such app, which can be downloaded from the website. There is also a web-based version (a demonstration that is available at the website) that has a pre-set template, which can help users get started a little faster. Dr. Holmes was interested in how World Wind was managed, and if there were any statistics on the user base, growth, or feedback mechanisms. Dr. Kim identified Patrick Hogan as the point of contact for statistics. NASA does track visitors to the website, and has seen a surging interest in the web-based version. World Wind is funded partly through reimbursable work from the defense community and others. Two NASA offices have an interest, Human Exploration and SMD, but they have an "internal adoption" problem. Dr. Holmes commented that it will be useful to have programmatic feedback to help WW survive.

NASA Cloud Computing Initiative

Dr. Mike Little presented a briefing on Cloud computing. Dr. Little is an Advanced Information Systems Technology (AIST) program manager, inside the Earth Science Technology Office at Goddard. The Earth Science Technology Office (ESTO) has been in existence for about 20 years, and has several programs that focus on instrument platforms. ESTO uses techniques to validate technologies that the science community would like to use within 5- 25 years, and works to familiarize the community with the available and maturing technologies, as well as tries to infuse them into the programs and projects. INVEST is an in-space validation program that flies technologies in space, mostly on cubesats, and on the International Space Station (ISS). ESTO's AIST addresses the generation of data through the exploitation of data, and is divided into three technology areas: Operations, Computational, and Data-Centric Technologies.

The Jet Propulsion Laboratory (JPL) has been using Cloud computing for the last six years. AIST has tried to use the JPL experience and associated Lessons Learned. The key problem is lack of experience with Cloud security by almost everyone. The Approval to Operate (ATO) process is complex and expensive (it cost $1M to implement FedRAMP, in one case). AIST is trying very hard to overcome barriers and understand the risks associated with Cloud computing, but the only way to understand it is to use it. AIST is demonstrating the value of the Amazon Web Service (AWS) to science PIs, the easiest mechanism possible to get scientists to try out Cloud computing. The task has been facilitated greatly by the diligent work of Dr. Ray O'Brien. AIST funds three major Cloud efforts: the AIST Managed Cloud Environment (AMCE); technology developments to enhance Cloud Computing usage such as the SAR Science Data Processing Foundry and ESD's EOSDIS; and ATO clearance for ARC-GIS Online Portal (SaaS for GIS tools).

AMCE is meant to support project teams in running research investigations; almost all the research projects are very heterogeneous (commercial, academic non-NASA, international). In many cases, the teams use students for research. Getting all of them credentials was an issue. The AMCE goal is to allow these multi-organization teams to share resources, and to try to get PIs access to Cloud computing within 24 hours, with the rest of the team coming along afterward. It is a pay as-you-go process; if you plan ahead and have appropriate funding, you can get a lot done. AMCE built a brokerage system to ensure that resource consumption is appropriately tagged. Mechanisms are put in place to freeze an activity if it goes over budget. AMCE also provides training, and tries to minimize management overhead; the objective is to understand how much it actually costs to run a virtual data center.

AMCE is aiming to decrease hardware/compute costs, and demonstrate a model for science users to use AWS in efficient ways. Currently, this is a low-risk system, and can be used to encourage people who really need supercomputer time to work through the HECC program to get the proper resources. The concept of operations was likened to running an apartment building: renting out space to scientists for particular capacities. A function of the elasticity of the computing environment is that it needs to be able to give people the appearance of unlimited capacity. When a user is not using it, s/he's not paying for it. Alarms are built in to alert users of overruns. AMCE program managers and PIs are notified before account lockdowns. AMCE also provides monitoring against hacking, and is training people to use "containers" to avoid errors and re-builds.

To transition project research to operations, AMCE is working to reduce the cost of refactoring to run in a NASA environment using accepted operating systems, libraries, and security measures, and providing a central authority for reviewing and approving image updates. SaaS helps to avoid re-installations. AMCE encourages use of source control and documentation tools, and seeks to improve the ability to conduct independent testing. One gap that is yet to be filled is the ability to be able to accurately estimate and project timing of Cloud costs.

Benefits of AMCE for the PI and the program include enabling quick and easy access to a lot of data, and collaboration between NASA PIs, university partners and industry. Four projects have been on-ramped thus far, and five more are in the pipeline. The program has been socialized at various seminars and presentations. When the AIST16 solicitation comes out, many projects will be using AMCE.

Dr. Little discussed the Synthetic Aperture Radar (SAR) Science Data Processing (SDP) Foundry at some length. The SAR SDP Foundry supports a number of satellites and UAVs that are generating SAR data, through a common tool to handle these data, which can be very extensive at centimeter-level resolution. A laptop can process one scene at a time, and it can take months to show three days of changes. Cloud computing is being used to process several thousand scenes over night, instead of 3-6 months. When the Napa Valley earthquake occurred, the Foundry was able to take data from a satellite overflight, and in one day delivered a damage map to the centimeter level to emergency management agencies. Once the process was established, many other SAR data users expressed interest, using standard products, and the same sources. Central Valley drought research, e.g., uses SAR data, as do land subsidence researchers in Virginia. The idea behind the Foundry is that once the products are run, users can use the exact same processing line on AWS. The idea is to see if the Foundry model is a good business model.

Dr. Feigelson commented that ROSES 2016 does not include AIST grants. Dr. Little explained that there would be an amendment to ROSES 16 to accommodate AIST. Dr. Holmes offered his congratulations for developing this partnership for exploratory research; He asked what was preventing outreach to non-ES programs. Dr. Little said he was in the process of spinning up outreach, indicating that the Chief Technologist for SMD is interested. He believed AIST could scale up to meet demand, and broadcast details through Brown Bag talks. Dr. Little was looking for an operational data center to partner with, and was having these discussions. Dr. Holmes asked if AIST would be amenable to demonstrations with the supercomputing group at ARC. He said he would be happy to defer to Drs. Mehrota and Lee. Dr. Mehrota felt AMCE might be great for "pleasingly parallel" runs for pipelines like Kepler and TESS.
Dr. Tino asked, regarding the need for culture change and training, if it were possible to embed software engineers in the project, and provide help desk support to make sure researchers have

access to help. Dr. Little reported that AMCE has an email helpdesk at present, and saw it as becoming more effective as the system is operationalized. Dr. Tino noted that multidisciplinary fields should make the best use of available resources, and so should understand infrastructure and software engineering. Dr. Little felt this was extremely relevant observation; does an atmospheric scientist need to know how to write machine-learning code? There needs to be a culture change to enable this collaboration. Dr. O'Brien said he was making these kinds of efforts in his program, bridging the gap between old and new. Dr. Holmes observed that data centers (serving, storing, analytics) at GSFC have found that existing cost models were not appropriate. Dr. Little replied that AIST doesn't store much data, and would thus rely on NGAP experience to answer this question. Dr. O'Brien felt that some NASA groups should take on one heavy lifting task on behalf of the entire Agency.

NASA/IPAC Extragalactic Database (NED)

Dr. Rick Ebert, a Senior Engineer at California Institute of Technology, provided a briefing on the NASA/IPAC Extragalactic Database (NED). NED is an archive of multi-wavelength data, the "Google of Galaxies," that enables querying and comparison of multiple sources, with key measurements from 22 NASA missions. Variables include distance, luminosity, and redshifts for 214 million astrophysical objects, in 7 billion data records. In user interfaces, NED has adopted Drupal, a common web application framework. NED has been getting into machine learning, and is running a small pilot project to apply machine learning to some internal processes (e.g., routing papers to the proper curators). The challenge to machine learning is in cleaning up the data to make a good training set. Database technology is always changing, so NED is always restructuring. Multiple versions are necessary in order to understand data prospecting and mining. There are Big Data implications for NED. Over 2016-19, NED will be growing tenfold, to contain 1B objects, with over 10B attributes. NED will have to refactor its database to improve scalability, extensibility and data-driven design. Data are often not properly referenced, and in response, NED has published, in collaboration with journals and centers, guidelines on best practices to correct this problem. Purely archival research using NED led to the discovery of a new class of galaxies: superluminous spirals..

The NED vision is to evolve the database, to enable powerful data processing queries. Statistical summaries of the NED contents are being developed to help users frame questions and use the data sets. Each day, NED serves 70K queries, and on average is cited daily in two new publications; NED has a literature citation-rate equivalent to a NASA Great Observatory. In response to questions from BDTF, Dr. Ebert remarked that a function NED would like to cease is its bibliographic search function. Currently NED is in the process of linking to the ADS, which will give users a broader interface. NED looks for the best information with future value for the community. Dr. Tino commented that NED seems to be one of the few groups that has a continuing budget for refactoring. Dr. Ebert noted that keeping this funding going is a mounting challenge. Dr. Feigelson asked if NED had any significant problems addressable by the BDTF. Dr. Ebert said that NED is at the terabyte level, and uses entry-level, data-center grade computational equipment. So far, NED has been leveraging existing technologies for parallel processing, but Dr. Ebert foresaw a day when data rates will present a formidable challenge.

Infrared Science Archive (IRSA)

Dr. Steve Groom presented information about the Infrared Science Archive (IRSA), which is chartered to curate the science data products from NASA's IR and submillimeter projects, including Spitzer, WISE, Planck, 2MASS, IRAS, and soon, SOFIA and IRTF. IRSA also accepts related enhanced data products from the community. In total, IRSA provides access to a petabyte of data, including all-sky coverage in 20 bands. The richness of this content attracts

many researchers, who have submitted 35.5 million queries and downloaded over 264 TB in the first 10 months of 2016. IRSA's impact on science is significant: about 10% of refereed astrophysics journal articles use data in IRSA's holdings. For the past several years, the majority of *Spitzer* publications have used archival data, demonstrating that archives have the potential to double the number of papers from a project.

Operational technologies have been applied in the last few years to accommodate the volume and flow of data. These address the need for users to quickly query very large tables, to visualize large and complex data sets, and to manipulate large volumes of pixel data. Fast queries and bulk queries were implemented in IRSA's Catalog Search Tool, which provides access to 100 billion table rows. The Firefly visualization package was developed by NASA's Infrared Processing and Analysis Center (IPAC) to link images, plots, and tables from multiple data sets. IRSA's web-accessible version of the WISE Coadder allows a user to tune parameters to make a custom mosaic of the sky.

IRSA users are carrying out science that requires them to contend with large volumes of imaging data. One example is the large-scale reprocessing of WISE images, which are difficult to move over the internet to a facility with adequate compute power. Another example is the search for interstellar features in Spitzer and 2MASS images, which has benefited from the participation of citizen scientists and machine learning.

In the era of Big Data, users will increasingly struggle to find the *right* data. The next generation of Data Discovery tools will need to meet this challenge. Users will also struggle to move large data sets from the archive to their home institutions, and will need archives to support science analysis in place. Possible approaches are partnering with NASA and other sponsors on developing a more analysis-friendly model; or using the Cloud as a convenient middle ground for shared purposes, a co-location services approach. The question is whether archive data volumes are compatible with Cloud cost models. Dr. Groom noted that the 2015 Senior Review gave low priority to technology pilot efforts, apparently not recognizing the importance of being able to do this sort of experimentation to serve users better.

IRSA responded to three general questions from the BDTF.
1) Dr. Groom described how IRSA plans for the future. Priorities come from the NASA IR projects (e.g. Spitzer, WISE) that deliver data to IRSA and from a dedicated User Panel that meets twice a year. The Archive Senior Review evaluates IRSA's plans on a 4-5 year timescale. Based on these inputs, IRSA sees a need to develop a next generation Data Discovery tool that will help users find the best data to meet their scientific needs, exploration tools that will better support multimission research, and data analysis capability at the archive.
2) IRSA would like to retire some old software tools that require maintenance. Software refresh activities must be scheduled between high priority activities and development of new functionality, as discussed with the User Panel.
3) IRSA has developed standard application program interfaces (APIs) that allow other archives (e.g. NED, MAST, HEASARC), Desktop tools (e.g. TopCat, ds9), and mission pipelines (e.g. PTF, Spitzer) to query IRSA. IRSA's APIs adhere to Virtual Observatory Protocols, as outlined by the NASA Astronomical Virtual Observatories (NAVO). To identify candidate data sets for integration with IRSA services, IRSA receives input from its User Panel, other members of the Astrophysics Data Centers Executive Council (ADEC), NASA missions/scientists, and the community.

NASA Exoplanet Archive (NExScI)

Dr. Rachel Akeson presented a description of the NASA Exoplanet Archive (NExScI), whose function is to gather data from the literature to maintain a comprehensive list of exoplanets. NExScI also curates some products from Kepler and other exoplanet transit surveys. The archive also predicts transits, and maintains a service called ExoFOP (Exoplanet Follow-up Observing Program). Current challenges are CPU resources and data complexity. Re: the CPU challenge, because demand is not constant, for example, when computing periodograms of light curves, NExScI is implementing a tiered support model in response. For larger jobs, the plan is to use Cloud computing, and for the largest jobs, to provide code to power users. Amazon, Google and Caltech are currently being evaluated as cost models. As to data complexity, NExScI deals with very diverse data, described differently, and with inconsistent terminology. For transit follow-ups, data is often time-critical. Solutions are to present data in multiple ways, thus for specific areas, NExScI uses focus tables, which allow users to configure and save preferred columns.

Over two-thirds of exoplanet discovery papers reference the Exoplanet Archive. Archive usage has been growing steadily since 2011, and was included in the 2015 Archive Review. The database is growing very quickly. Because the Exoplanet Archive is relatively young, there are not yet any tools NExScI wishes to cease supporting, but it is concerned about scaling its current tools. While current tools are still relevant, the archive will consult its Users Group when this changes. NExScI provides a weekly update of nasa.exoplanets.gov, and also houses links to IRSA and the Mikulski Archives for Space Telescopes (MAST). Overall, NExScI is prepared for the expected growth in the next 10-15 years. Dr. Walker asked if less relevant data were being removed from the archive. Dr. Akeson felt it unlikely that the data would become less relevant. The community has asked the archive to host exoplanet-related data from non-NASA sources, which has a cost associated with it; NExScI may have to consider off-loading it in the future. Dr. Holmes asked if the archive were capturing only light curves from Kepler. Dr. Akeson explained that MAST receives the pixel data, while NExScI gets the higher-level mission data products. Dr. Tino asked if NExScI allows users to cache regular queries. The response was that it allows them to cache table configurations. Dr. Tino observed that some users like to move data sets back and forth, and was curious to see if this behavior dies off over time. Dr. Akeson felt that people were not entrenched quite yet. As the missions get bigger, standardization will probably increase. It's not just a data-gathering task; there will need to be some science input.

Dr. Feigelson asked Dr. Ebert why NED was concerned about incorporating non-NASA data, asking: Does NED seek to provide an integrated approach to all galaxy data in the current era? Dr. Ebert felt the fundamental question is which of these data sets has the best potential for supporting future missions. Dr. Holmes said it was important to note that NASA is funding PanSTARRS to some degree. Dr. Groom noted IRSA does accept non-NASA data, and that its purpose is to be a repository and a research resource for NASA missions. IRSA accepts these data with the full knowledge of NASA Headquarters.

Public Comment

No comments were noted.

BDTF Study Topics

The BDTF briefly discussed the day's presentations. Dr. Beebe didn't think NASA could use anything other than a pay-as-you-go approach to Cloud computing. On the other hand, she noted how many queries are generally received on how to write a successful proposal; researchers typically don't have the money to answer their questions through Cloud computing,

which could become an exclusive service if not carefully handled. Dr. Holmes noted two potential uses of Cloud computing for NASA: data processing, modeling, and analytics; and data storage and retrieval. Commercial storage is currently too costly for NASA. A Science Cloud at GSFC could be an answer to that. Using pay-as-you-go to do processing, however, can be a cost item in a proposal. Dr. Tino commented that potential egress bandwidth is a risk. Dr. Mehrota noted that people often use their own funds to pay for egress from AWS. Dr. Tino felt that the onus could be on the group to get data into storage, after which a hands-off approach could be used. It would be possible to build in Cloud storage costs as fixed costs.

Discussion and individual reports on study topics
Dr. Holmes raised the issue of the need to streamline the process for modeling workflows in areas such as climatology, star formation, and atmospheric dynamics. Dr. Walker said he had been discussing the topic with Dr. Kinter, determining the difference between computing for plasma physics vs. Earth Sciences. There are community models that can be used for Earth Science. In plasma physics, the models are purely experimental. The closest design for plasma physics is a model that is being deployed for space weather prediction. Here it makes more sense to run the model, look at selected products off-line, and proceed. There is more than one way to do it. Plasma physics is not at the point where we have standard models that everyone can use. Dr. Walker asked a bunch of simulators about using Pleaides in such a manner, and got a negative response. The community really needs more information. Dr. Tino felt that each community has very specific expectations, some of which are amenable to the infrastructure and some not. Dr. Beebe noted that there are four different Mars models, all competitive, and all evolving rapidly. Researchers have been told to encourage transparency of the models to serve the community. Dr. Walker felt it useful to refine the work plan to reflect the diversity of approaches. Dr. Tino suggested classifying what approach each community expects; as the BDTF shouldn't get bogged down in recommending types of models; this also brings in the question of whether NASA can afford to serve everyone.

Dr. Beebe commented, re: data discovery, that one goal is to identify areas that need improvement or expansion. She reported having circulated a request to BDTF to identify users inside and outside of NASA (e.g., NOAA, NCAR), and has also polled faculty as to their personal use. In the next 3-4 months, she expected to be able to start to look through the areas, develop a set of data characteristics that will serve user needs, and identify individual archives within each SMD division. She expected there would be a recommendation for a future study.

To improve data analysis methodologies, Dr. Hurlburt felt that NASA should devise some plan and formulate strategies for incorporating these methodologies into the NASA community. Dr. Feigelson observed that ESD has incredible programs for analysis and visualization, based in part on a $25/M year program (AIST). Is the same fraction of budget allocated to each SMD division? NASA is trying to cope with large amounts of data and programmatic challenges. Astrophysics doesn't have Big Data, per se, but there is a vast amount of computing going on. Historically, all the big theory work is done by NSF using the kinds of resources described by Larry Smarr. In the Astrophysics discipline of simulating star formation, globular cluster problems are also carried out by NSF and in Europe. The rationale for JWST is galaxy formation; the Earth Science community perceives NASA's role as building Earth-observing satellites and curating the data. But the ES community also feels that NASA's job is to do the model. What if NASA offered Pleaides to process data from HST and JWST? If NASA is looking for new opportunities, Astrophysics is a ripe field.

Dr. Mehrota thought perhaps the recommendation should be to expand HPC to include more people who use NASA data. Dr. Tino noted that it comes back to what NASA's charter is- where does it end? Is it just collecting data? This gets right back to how people expect to interact with data. Dr. Holmes felt the discussion should also include NOAA and USGS. Dr. Feigelson suggested querying David Spergel as to what the Simons Foundation was seeking. Dr. Beebe suggested taxing the missions for HPC usage. Dr. Feigelson noted that there are observers vs. theorists, and users vs. interpreters in Astrophysics. In Earth Science, however, the theorists are as deeply involved as the observers. Dr. Beebe noted that Astrophysics typically does not have paying customers. Dr. Tino noted that a finding on needing more funding was not particularly strategic. Dr. Holmes hoped BDTF could at least influence prioritization of existing resources. He likened the problem to the "new bookcase law;" a new bookcase is immediately filled up. Dr. Mehrota said that time allocation for Pleiades is always tight. Dr. Tino felt that there seemed to be a big focus on hardware refresh but not science software refresh; this approach does not gain the efficiency of infrastructure. Dr. Beebe suggested one approach could be holding some money back for end-of-mission data processing. Dr. Tino agreed that a proposal should make the commitment upfront. Dr. Tino asked about the average age of codes. Dr. Mehrota said that it varies; there are very few third-party cases, and codes are always developing, but the software architecture is not changing. Dr. Feigelson thought there was too much emphasis on software implementation and deliverables, but not enough discussion about the quality of analysis methods underlying the software. He cited a satellite mission center that froze the science analysis software a dozen years ago, not incorporating more sensitive methods being developed by other experts. Another satellite team continuously improved its sophisticated analysis methods, but without oversight by external astronomical or statistical experts. Drs. Feigelson and Hurlburt suggest that algorithms and methods, as well as software implementations, be explicitly reviewed and funded within NASA missions.

The BDTF discussed the case for adopting server-side analytics. Dr. Holmes reported trying to articulate some common architecture, and maybe provide some case studies. Drs. Feigelson, Tino, Hurlburt and Holmes would be working on the topic, with plans to engage GSFC, GISS, and IRAS. Dr. Hurlburt felt that iPython and Jupiter were two good ideas. Dr. Tino preferred to avoid the use of the word "architecture," as there many ways to co-locate data and compute; BDTF should present a variety of options.

Dr. Holmes set a goal for the next meeting for producing outlines of the topic studies. Dr. Smith encouraged more talk about cross-disciplinary skill sets for inclusion in the study topics. Dr. Feigelson was impressed by how the Earth observation groups provided their services and thought they could promulgate expertise by training their counterparts in the other SMD divisions. Dr. Mehrota added that BDTF should consider discovery of tools, algorithms, and models. Dr. Tino thought cross-disciplinary expertise should be another study topic.

Potential F&R:
Given that the Science Committee wants the BDTF to relook at the FIONA platform, Dr. Holmes postponed this finding pending further research.

Dr. Tino suggested a finding on prioritizing software maintenance costs across a mission lifetime to keep pace with changing technology (to be written with Dr. Beebe).

Dr. Holmes suggested a demonstration of pipeline processing on Amazon for the exoplanet program. Dr. Tino felt this might be better as a good proof point, not a finding.

Dr. Hurlburt felt that open access to archival data has more than doubled the scientific productivity of many missions.

Dr. Holmes said he would work on consolidating responses to the three BDTF questions, and distill them for the next meeting, to see if there are some big picture results. He thought there was already good strategic evidence of the value of the archives. Dr. Tino found it interesting to note that archive usage increased dramatically when the archives started offering more interactive services, curation, and access using programmatic APIs. There seem to be two major questions-How do I get a specific outcome and; What is the API that tells you where to get the data you want?

Dr. Feigelson felt that NASA was excellent compared to other agencies and countries, and that it has the most capable Astrophysics archive system that links published literature with data.

September 30, 2016

Draft Findings and Recommendations
The Committee spent the morning refining five findings regarding the OCIO AIST task of doing scientific processes on Cloud computing services; a demonstration for Kepler/TESS on pipeline processing in lieu of Pleaides; adjusting the mission contracting process to enable a portion of budget for software processing and software maturation; APIs; and scientific use of archival data/papers.

Finding on "Observing the Archives"
There is increased use of archival data as reflected by the increasing number of published papers based on archival data, which doubles the scientific productivity of the missions. Paper/data source tracking is done by "people in the loop." Dr. Feigelson noted that Astrophysics journals have a LATEX macro for the author to indicate all the astronomical objects and observatories to automate. Authors use them unevenly rather than obligatorily, but there has been progress. Dr. Hurlburt added that there were similar efforts in Heliophysics. Dr. Holmes suggested directing this finding to SMD as an important statement for Headquarters to reflect on. An Ames participant commented that machine-learning algorithms are being used at Ames to do similar annotating. Dr. Lee mentioned that NASA does have R&D work in this direction, discovering data sets, tools, workflow, adding that IBM's Watson program invested 25 people for 5 years, full-time. NASA doesn't have the resources for such a large project. Dr. Oza thought a smaller version might be possible; as is done in Earth Science, one could take scripts and turn them into a part of a workflow, with raw data to publication represented.
The BDTF generally concurred on the finding.

Finding on "OCIO/AIST collaboration on Cloud Computing"
The finding essentially expresses approval of the recent launch of four AIST projects into the Cloud as a major milestone. Dr. Holmes wanted to impress that BDTF is especially encouraged that the two groups are finally collaborating. Dr. Walker said he would like to see a "Consumer Report" on scientific processing in the Cloud.
BDTF concurred on the finding.

Finding on "Kepler/TESS should collaborate with AIST to demonstrate exoplanet data processing in a Cloud computing setting rather than on the Pleiades supercomputer"
Dr. Beebe commented that the two missions seem quite different, as Kepler was sit-and-stare, and absolutely parallel, while TESS is moving/sweeping, with different calibrations and

numbers of points changing throughout the mission; the latter is a much more difficult problem. An Ames staff member noted that a Kepler demo would not be trivial, resource-wise, and was unsure how straightforward it would be to prepare the Kepler data for the demonstration. Dr. Tino felt it was essentially just a scheduling matter, shipping out jobs. Drs. Holmes and Feigelson agreed to follow up and determine how difficult it would be. Dr. Feigelson felt it was an easy enough problem, fairly modest, and a nice demonstration for another branch outside Earth Sciences. Dr. Holmes characterized the recommendation as a request to collaborate with ongoing work; a demonstration in theory should be a small effort. The point is that Pleaides is oversubscribed with some tasks that might be better/more efficiently performed in the Cloud. Dr. Tino recommended adding some words about structure, and that "cheap compute" could work just as well. Dr. Smith cautioned against recommending an unfunded mandate. Dr. Tino felt that SMD should look across divisions to see where similar demonstrations can be made. Dr. Holmes felt the cost shouldn't be more than tens of thousands, to use successful, existing infrastructure. Hurlburt- point is to relieve pressure on Pleiades. Dr. Lee cautioned against thinking that NASA HPC is a free resource; it is not. Dr. Holmes felt BDTF should take a deeper look at the allocation models and suitability of the problem, and follow up with Dr. Lee.

Dr. Smith expressed concern about how such a recommendation will be interpreted: i.e. don't pay Spitzer Guest Observers, and use this money for the demo. She felt that examining which missions are suited to Cloud computing was a more acceptable idea. Dr. Holmes viewed the recommendation as a small step with long-term gain, and an important technical step with strategic implications. Dr. Feigelson agreed with Dr. Holmes on this issue. Dr. Tino also felt cautious about introducing an unfunded mandate.
BDTF concurred on the finding.

Potential finding on "Mission software evolution."
The Task Force discussed the risks of the current mission paradigm, recognizing that mission software must maintain pace with industry standards, and in some cases, must be able to re-architect mission software where appropriate. Dr. Tino felt it useful to separate analysis software from pipeline software; it is a risk, but what's the value of the science data if you can't understand it in 30 years? Dr. Beebe recommended archiving raw data, and including algorithms that convert the raw data to calibrated and derived data. She cited a classic case wherein it took a horrendous effort to "redden" Mars; the original data was interpreted as blue in color. Dr. Walker felt the biggest issue was to preserve the algorithms. Dr. Feigelson thought the distinction between algorithms and software was crucial. Dr. Tino noted that there must be an understanding of what the algorithm is trying to achieve and the platform used to implement it. Dr. Beebe commented that the issue is that we have lousy temporal and spatial coverage in our data sets. Dr. Holmes asked: How does a program manager allocate resources for 10-20 years down the road when we can't predict what the practices will be? Dr. Tino noted that in business, one can develop features over time, as there's no good fudge factor. There does seem to be an ability to predict hardware refreshes, however, and there is a similar way to do this for software. If the data tail drops off in 20 years, what good is it?

Dr. George Helou offered as an example, 2MASS, which needs to be kept alive because most astronomers rely on it; the art is in finding ways to encode the important information, in documentation perhaps, and in making products that minimize future software runs. Dr. Tino observed that if the demand for older sets of data rises, the system built to support it is not going to be sufficient; NASA needs to encourage scientists to think longer term. Dr. Holmes offered to do more homework on the subject and obtain input on Senior Review and mission resource profiles. He tabled the finding for the interim, but agreed to mention to the Science

Committee that the finding was a work in progress. Dr. Feigelson suggested that NASA missions, in advance of launch, should write the algorithms and have engineers could vet them. He added, noting that statisticians are often viewed as too costly, suggested recommending that missions should include *ab initio*, software reviews, software scientists and mathematicians.

Dr. Oza introduced a brief discussion of the NASA Big Data Working Group (BDWG), which was stood up out of OCIO (all NASA). The Group meets every 6 months, and holds a telecom every month, and discusses Big Data projects, new initiatives, programmatic recommendations, etc. The group has 35-40 members. Dr. Oza felt that BDTF recommendations could be passed easily to the WG. Dr. Holmes requested that a BDWG briefing be put on the next meeting agenda.

<u>The Infrared Processing and Analysis Center (IPAC) in the Big Data Era</u>
Dr. George Helou presented details of the Infrared Processing and Analysis Center (IPAC) at Caltech. The guiding principle in Astrophysics archiving is enabling researchers to interact with data. IPAC is a science operations center and hosts data from Astrophysics and Planetary missions, with a special emphasis on IR and submillimeter astronomy and exoplanet science. IPAC also supports NSF and privately-funded projects. IPAC started with the Infrared Astronomical Satellite (IRAS) mission and then produced the Two Micron All Sky Survey (2MASS), which at the time was thought to be too data-intensive for in-house operations. NED was brought on in the 1990s, as well as ISO, an ESA-led mission. Spitzer has been a major IPAC mission, as have Herschel, Planck, WISE, and the NASA Exoplanet Science Institute activities. In the last five years, IPAC has diversified and built on expertise to support ground-based surveys such as the Palomar Transient Factory and the new Zwicky Transient Facility. IPAC supports asteroid searches via Near-Earth Object (NEO)-WISE, and is preparing to support the ESA Euclid and NASA Wide Field Infrared Survey Telescope (WFIRST) missions. IPAC has unique expertise in Astrophysics, systems engineering and software development, and produces 200 refereed papers per year. The staff stays up-to-date on modern technology, including big data techniques, via conferences, pilot projects, hiring new expertise, and encouraging existing staff to take advantage of online classes. Overall, IPAC has learned to rely more on reusable architecture, and not just on reusable code. Open-source Firefly components are prepared to support missions of the future, such as the Large Synoptic Survey Telescope (LSST). IPAC strengths and capabilities lie in the expertise and experience of its staff, its experience in supporting NASA missions and ESA-NASA partnerships, and its three data centers with thousands of servers and over 12 PB of spinning disk. Five terabytes (TB) of data are downloaded every week. IPAC's analysis is that even though cloud-computing can support special "ephemeral computing" needs, locally-hosted datacenters are still cost-effective, and expects that this will still be the case for the next five years. IPAC is participating in the NSF Pacific Research Platform (PRP), which aims to operate at a speed of 10-100Gb/s end-to-end. Currently, hardware and software components are being deployed to plug into the platform to see what the gains are. IPAC does not receive resources to do this; it is doing this to be able to support the ESA-led Euclid mission, which will require moving data quickly and ensuring connectivity. PRP is a multidisciplinary enterprise that is valuable for new ideas and techniques.

Even if the typical Astrophysics data set is not huge, the data sets are growing fast. The vision is to enable discovery beyond the design drivers of archives and tools, similar to observatories discovering the unexpected (superluminous spiral galaxies), and enabling researchers to interact meaningfully with petascale data sets. IPAC is developing techniques, e.g., for LSST Science User Interface and Tools, and for IRSA, to enable research that is different from that envisioned by the original mission design drivers.

Big Data elements for Astrophysics are manifested in the exploding data volumes, huge database tables, diversity of data types, and dynamic holdings. Examples of practical applications for cloud-computing include the Sagan workshop's use of Amazon Cloud services (AWS was used for short-term intensive computing). Montage, a portable toolkit for creating science grade astronomy image mosaics, has been set up in the Cloud, and is associated with 290 literature citations. A Herschel virtual machine service was also set up in an IPAC-hosted "Virtual Cloud", providing access for many users to complex and resource-intensive Herschel analysis software, and resulting in several discoveries of brown dwarf stars. The future holds more surveys from both ground and space: LSST, WFIRST, Euclid, and the Zwicky Transient Facility.

To meet the data challenges of science opportunities arising from joint analysis of two or more individual surveys, a study has been initiated to target specific science goals, identify best uses of available assets, and use of virtual machines. Processing tiers may be one approach to meet the challenges of joint processing of large surveys, to maximize science returns at fixed cost.

Discussion

Dr. Tino commented that big data analytics will have to span the current stovepiping common at NASA. He asked whether APIs were getting more uptake, and yielding a better return on investment by increasing opportunities for interfaces. Dr. David Imel said that IPAC was certainly seeing an increase, and that the Virtual Observatory protocols have made a distinct difference. Dr. Helou added that the protocols allow people to see much more quickly the data that is useful to them, and to download them. He felt that IPAC could still offer better tools to explore the statistical nature of a sample, to understand where the sources lie, and where the outliers and interesting concentrations reside. There is also a time domain aspect; each source may have 10-1000 visits. If you want to find sources that have specific time domain signatures, you can't apply simple search algorithms. IPAC would like to support a way to do a proxy search and then refine it.

Dr. Feigelson asked whether there was a true scientific need for a pixel-by-pixel comparison vs. catalogue comparison. Dr. Helou agreed that some images be dealt with as catalogue entries; making an effort to minimize need for pixel comparisons; however, there is a need for forward modeling of some objects that have very different redshifts, for example, for the purpose of disentangling apparent image overlaps. Dr. Holmes asked whether architectures needed to be preserved for reuse from mission to mission. Dr. Helou said IPAC puts together documents that describe both the algorithms and data processing, which accompany the data products. Dr. Imel felt that the archive doesn't need to be reconstituted, one just needs to know how it was built. Dr. Holmes felt that holding reviews every 4 to 5 years was totally inadequate to keep pace with new tools, and thought BDTF may need to recommend to Headquarters an increase in the frequency of programmatic reviews for the archives. Dr. Helou said the archives do have access to the Planning, Programming, Budget and Execution (PPBE) process every year, and thought the reviews were about right in cadence. Dr. Feigelson asked how NASA archives decide, programmatically, what to archive from non-NASA surveys. Dr. Helou said this was based on reimbursable services, with outside funding. For the MAST archive re; PanSTARRS, he believed NASA was using internal resources, because the data sets were judged as scientifically valuable. Dr. Holmes reiterated that PanSTARRS does get NASA funding, and that archival inclusion at NASA is generally decided on a case-by-case basis, with scientific utility of the data as the paramount question.

Study topics

Dr. Holmes thought BDTF should consider a one-day meeting in December 2017 (AGU) to close out a final report and finalize steps of the studies, using early summer 2017 to polish up a rough draft of the report. The next two-day meeting is scheduled in Washington, DC in February 2017, with briefings from Tsengdar Lee, Kevin Murphy, Jeffrey Hayes, and Hashima Hasan, a conclusion of the Big Hubs discussion, a briefing DOE's exoscale project, and an update on AIST's Cloud computing project. Dr. Beebe suggested a briefing from the Chief Engineer at PDS, who is working in medical sciences.

Public Comment
No comments were noted.

Lawrence Berkeley National Laboratory: DOE DMZ
Dr. Eli Dart, network engineer of at DOE's Energy Sciences Net (ESNet), presented a briefing entitled "The Science DMZ Design Pattern." ESNet is a facility that connects the Lawrence Berkeley National Laboratory complex to the scientific and public Internet. ESNet and NASA have collaborated on this facility for many years. It is a small organization of about 40 people, trying to make location irrelevant for scientific purposes. ESNet has a national-scale 100Gb/s backbone, and also has connections with Europe.

Networks are essential to data-intensive science. "Data is growing exponentially but people are not" is an apt description of the problem. Wide Area Networks (WAN) have not been an effective tool for scientists; ESNet trying to remedy the problem by facilitating data flow from experiment to analysis, and between facilities. Correctness, consistency, and performance are the most important factors, in order of importance. Transmission Control Protocol (TCP) is ubiquitous and fragile; this is how the host sees the network. TCP is timid; packet loss is interpreted as congestion, and loss in conjunction with latency is a performance killer. Practically speaking, it's easier to support TCP than to fix it, by having sufficient bandwidth to avoid congestion. A solution must be easy to adopt and to secure. In traditional network security, a DMZ houses and connects external-facing services, creating a clean separation from the internal network. ESNet is doing this for science, hence a Science DMZ, which is designed to move data in and out, period.

The minimal configuration for such a network is composed of a WAN connected to a border router connected to a Science DMZ Switch/Router, with appropriately located security control points. ESNet can give a project its own connection to the DMZ, or can use campus dark fiber routes to connect to the border router, creating small DMZs in several locations. In the case of supercomputing, data transfer nodes can be added to a parallel file system, which can then ingest from the data transmission network (DTN) and immediately launch a compute job.

Common threads of ESNet include the accommodation of TCP, and the ability to test and verify data via perfSONAR nodes. PerfSONAR is a widely deployed testing infrastructure that monitors performance. ESNet uses dedicated systems (data transfer nodes with limited ability); this is also important to performance. Science DMZ security works by segmenting risk into different enclaves and applying security in isolation to different areas, using the principle of least freedom (removing degrees of freedom).
NSF has funded many science DMZs; NIH has incorporated them, as has the USDA, Australia, New Zealand, and the UK. Others can build upon these deployments. The petascale data movement is now possible outside the LHC experiments, and the PRP is being built on top of the growing infrastructure. Science data portals, with large repositories of science data, haven't changed in 15 years. To first order the data portals are web pages, which is not suitable for

petascale data. Legacy portal design would be very difficult to change without architectural change. The next generation portal leverages the science DMZ, and will only hand out data based on a properly authenticated request.

The academic Globus project (Globus.org) is hosting a petascale DTN project, based on collaboration between the national labs; it has some work to do yet, but some labs are reaching 20 Gb/s.

Dr. Hurlburt asked whether the DMZ requires a clean dedicated network. Dr. Dart explained that it requires a loss-free IP layer, thus the public Internet not suited to it. That's why there are mission networks. Dr. Holmes asked how NASA could persuade science management to work on networking. Dr. Dart described the data as being outside a firewall but not outside a data filter. The strength in an enterprise firewall is in analyzing a large number of data packets (emails, image renderings). Dr. Dart suggested telling managers in a way they understand, and get them to do a risk analysis based on actual deployed tools. There is a critical sociological component; the organization has to recognize that they're causing mission problems by continually putting obstacles in place.

Dr. Dart gave a few more details. LHC is global in scale, while a petascale DTN is being developed in California, Illinois, and Tennessee. Larry Smarr is adding to the PRP fold. The idea is very scalable and PRP is at the forefront. What's important is the architecture, not the instantiation. Ideally, the Science DMZs should be connected such that the whole is greater than the sum of its parts. The community needs to educate people about the tools available and explain to them how it makes their work easier. Dr. Dart said he would be happy to talk to Ames staff, and would like to learn more about NASA cyber policy to identify barriers.

<u>Wrap-up</u>
Dr. Holmes closed the meeting by reiterating the three big items for BDTF's next steps: revisit Data Hub contacts look for opportunities for NASA scientists to engage; work on study topics; summarize the responses to the three questions and condense them into proper statements at the next meeting.

Appendix A
Attendees

Ad Hoc Big Data Task Force Members
Charles P. Holmes, ***Chair***, Big Data Task Force
Reta Beebe, New Mexico State University
Dr. Eric Feigelson, Pennsylvania State University (webex)
Neal Hurlburt, Lockheed Martin
Clayton Tino, Virtustream, Inc.
Raymond Walker, University of California at Los Angeles
Erin Smith, ***Executive Secretary***, NASA HQ

NASA Attendees
Jennifer Dungan, NASA ARC
Sangrum Ganguly, NASA ARC
Kim Hines, NASA ARC
Jon Jenkins, NASA ARC
Randolph Kim, NASA ARC
Tsengdar Lee, NASA HQ
Pamela Marcum, NASA ARC
Piyush Mehrota, NASA ARC
Rama Nemani, NASA ARC
Ray O'Brien, NASA
Nikunj Oza, NASA ARC
Gerald Smith, NASA HQ
Petr Votava, NASA ARC

Non-NASA Attendees
Rachel Akeson, Caltech-IPAC/NExScI
David Bell, USRA
Eli Dart, ESNet/LBNL
Rick Ebert, Caltech-IPAC/NED
Steve Groom, Caltech-IPAC/IRSA
George Helou, Caltech-IPAC
David Imel, Caltech
Amy Reis, Ingenicomm
Joan Zimmermann, Ingenicomm

Appendix B
Membership Roster

	Name	***Role on Committee***
1	Dr. Charles Holmes Retired - NASA	Chair
2	Dr. Clayton Tino Virtustream, Inc.	Aeronautics
3	Dr. James Kinter, III George Mason University	Earth Science
4	Dr. Neal Hurlburt Lockheed Martin	Heliophysics
5	Dr. Raymond Walker University of California, Los Angeles	Heliophysics
6	Dr. Eric Feigelson Pennsylvania State University	Astrophysics
7	Dr. Reta Beebe New Mexico State University	Planetary Science
8	Dr. Ashok Srivastava Verizon	Cyberinfrastructure

Appendix C
Presentations

1. Welcome to Ames Research Center; *Tom Edwards*
2. NASA Big Data Challenge: Ames Perspective; *Piyush Mehrota*
3. NASA Ames Data Sciences Group; *Nikunj Oza*
4. NASA Earth Exchange (NEX); *Rama Nemani, Petr Votava, Sangram Ganguly*
5. Supercomputing in the Age of Superearths, Earths, and Exoplanetary Systems; *Jon Jenkins*
6. Enabling NASA's Use of Cloud Software-as-a-Service; *Ray O'Brien*
7. NASA World Wind; *Randolph Kim*
8. NASA Cloud Computing Initiative; *Mike Little*
9. NASA IPAC Extragalactic Database (NED); *Rick Ebert*
10. Infrared Science Archive (IRSA); *Steve Groom*
11. NASA Exoplanet Archive (NExScI); *Rachel Akeson*
12. Infrared Processing and Analysis Center; *George Helou*
13. Lawrence Berkeley National Laboratory DOE DMZ; *Eli Dart*

Appendix D
Agenda

Ad Hoc Big Data Task Force
of the
NASA Advisory Council Science Committee

September 28-30, 2016

NASA Ames Research Center
Building 152, Rm 116/117

Agenda
(Pacific Daylight Savings Time)

Wednesday, September 28

9:00 – 9:45	Opening Remarks / Introduction	ARC Management Dr. Erin Smith Dr. Charles Holmes
9:45 – 10:15	NASA Big Data Challenges: Ames Perspective	Dr. Piyush Mehrotra
10:15– 10:45	NASA Ames Data Sciences Group	Dr. Nikunj Oza
10:45 – 11:00	***BREAK***	
11:00 – 12:00	NASA Earth Exchange: Helping Scientists Tackle Big Data	Dr. Rama Nemani Dr. Petr Votava Dr. Sangram Ganguly
12:00 – 1:00	***LUNCH TALK: SOFIA Science Highlights***	Dr. Pamela Marcum
1:00 – 1:05	Public Comment	
1:05 – 1:35	Supercomputing in the Age of Discovering Superearths, Earths, and Exoplanetary Systems	Dr. Jon Jenkins
1:35 – 2:30	Member Reports/Discussion	
2:30	***ADJOURN FOR DAY 1***	

Thursday, September 29

9:00 – 9:30	Enabling NASA's use of Cloud Software-as-a-Service	Raymond O'Brien
9:30 – 10:00	NASA World Wind	Dr. Randolph Kim
10:00 – 10:10	***BREAK***	

10:10 – 11:10	NASA Cloud Computing Initiative	Dr. Mike Little
11:10 – 11:30	NASA/IPAC Extragalactic Database (NED)	Dr. Rick Ebert
11:30 – 11:50	Infrared Science Archive (IRSA)	Dr. Steve Groom
11:50 – 12:10	NASA Exoplanet Archive (NExScI)	Dr. Rachel Akeson
12:10 – 1:00	***LUNCH***	
1:00 – 1:05	Public Comment	
1:05 – 1:50	BDTF Study Topics	
1:50 – 2:30	Discussion	
2:30	***ADJOURN FOR DAY 2***	

Friday, September 30

9:00 – 10:00	Draft Findings/ Recommendations	
10:00 – 10:45	BDTF Study Topics	
10:45 – 11:00	***BREAK***	
11:00 – 11:30	The Infrared Processing and Analysis Center (IPAC) in the Big Data Era	Dr. George Helou
11:30 – 12:00	Discussion	
12:00 – 1:00	***LUNCH***	
1:00 – 1:05	Public Comment	
1:05 – 2:05	Lawrence Berkeley National Lab: Department of Energy ESnet	Eli Dart
2:05 – 4:15	Complete Findings/ Recommendations	
4:15 – 5:00	Final Discussions, Next Meeting, and Conclusion	
5:00	***ADJOURN***	

Dial-In and WebEx Information

For entire meeting September 28-30, 2016

Dial-In (audio): Dial the USA toll-free conference call number 877– 601–6603 or toll number 1–517–319– 9533 and then enter the numeric participant passcode 4718658 . You must use a touch-tone phone to participate in this meeting.

WebEx (view presentations online): The web link is https://nasa.webex.com, the meeting number is 990 210 984, and the password is BDTFmtg#3 .

**** All times are Pacific Daylight Savings Time ****