OCIO

Mission to Mars

A Study in Data Volume and Storage

18 April 2012

Stephen O'Hearn

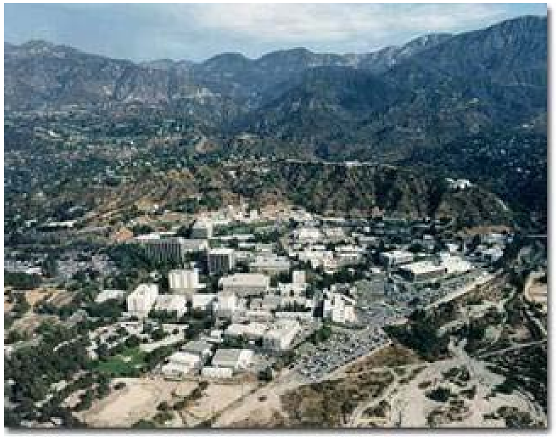
Jet Propulsion Laboratory – California Institute of Technology

Agenda

- What is JPL and what do we do – just in case you didn’t know…
- Data Types: What are they? How do they matter during the mission?
- Data Growth: Why is it growing? Who or what is driving it?
- Data Use: Where does it get used? Who by?
- Data Retrieval: Finding it afterwards? Or not.
- Data Recovery: Uh oh!
- Lessons Learned

What, or better who, is JPL?

- JPL, or the Jet Propulsion Laboratory, is a NASA Federally Funded Research and Development Center
- Managed by Caltech for NASA
- Responsible for robotic exploration of the Solar System
- $1.7B contract per year, ~ 5,000 employees; 177 acre facility located in Pasadena, CA, with 670K ft^2 of office space and 900K ft^2 of labs

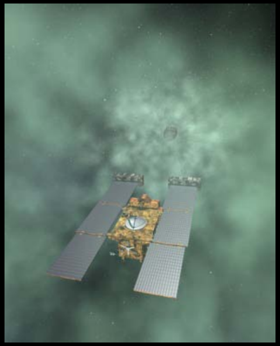


What JPL Does?

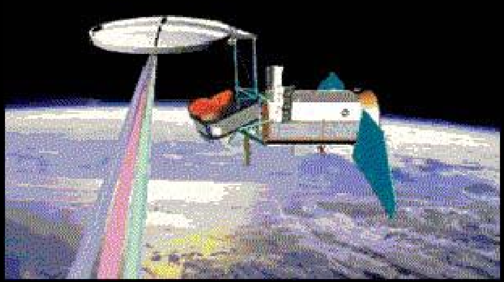
- Manages the worldwide Deep Space Network (DSN):
  - 3 Locations - Goldstone CA, Madrid Spain, Canberra Australia
  - Spacecraft Command & Control - recording scientific data
- 50+ years experience in spacecraft design, production, and operation
- JPL spacecraft have visited all the planets in our Solar System except for Pluto! But…
  - New Horiozons arrives at Pluto in 2015
- Keystone missions like:
  - Explorer 1
  - Ranger, Surveyor, Mariner (early Lunar and Inner Planet)
  - Viking (Mars lander)
  - Voyager (Grand Tour)
  - Galileo and Cassini (Jupiter and Saturn with probes)
  - Mars Rovers: Sojourner, Opportunity, Spirit, Curiosity

Near Term Mars & Solar System Exploration Events

EPOXI
Comet
Flyby
Nov. 2010

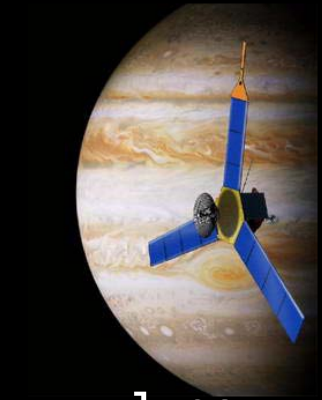


Stardust-
NExT
Comet Flyby
Feb. 2011

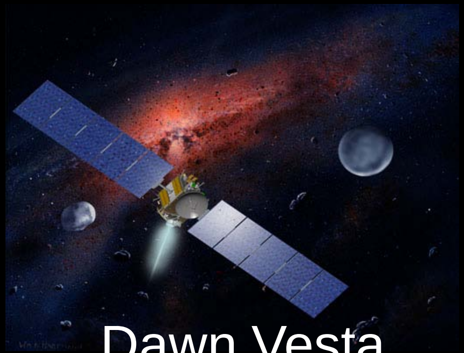


Aquarius

Jun. 2011



Juno
Aug. 2011



Dawn Vesta
Arrival
Aug. 2011
(Ceres, February
2015)

GRAIL
Sep. 2011

Mars Science
Laboratory
Nov. 2011 /
Aug. 2012

NuSTAR
Mar. 2012

Background – Data Types

- Engineering – data that helps build it and keeps it running:
  - Developmental and test data for the spacecraft, its instruments, and supporting ground systems
  - Huge volumes prior to launch, often predicted unsuccessfully, then steady state (spacecraft/rover operations telemetry)
- Science – why we sent it there in the first place:
  - The product of the instruments/experiments
    - Volume and rates driven by data resolution and power (electrical and RF – transmission back to Earth) budgets
  - Volumes can be enormous especially after processing and science product production
    - Reprocessing of a prior mission’s data with new techniques
  - Extended spacecraft operations adds even more, often years’ worth

Growth – There and Back Again

- What is leading it? Appetite and production – water, life, evidence of same:
  – Instruments/sensors and power for them allow fine resolution, equals larger data volume and more importantly higher data rates
  – New techniques allow new science to be extracted from “old” data
  – Some things just won’t die: Voyager: 35 years, Mars Rovers: 8 years vs. 90-days
- How is it possible? Design and leverage:
  – Spacecraft power sources, communications subsystem and better signal encoding, antennas systems there and here on Earth, and highly sensitive receivers in the Deep Space Network
  – Also using <u>other</u> spacecraft as data relays

Growth – Mission Data Rates

- Data rate values and impact of distance to Earth:
  - Voyager*
    - 115.2 Kbs at Jupiter, 44.8 Kbs at Saturn, 21 Kbs at Uranus (required a 3-antenna array)
  - Galileo* at Jupiter
    - 134 Kbs (design value, significantly reduced because of High Gain Antenna failure)
  - Cassini* at Saturn
    - 56.3 Kbs
  - Mars Odyssey
    - 256 Kbs
  - Mars Reconnaissance Orbiter (MRO)
    - 6 Mbs

* Radio-isotope powered instead of solar

Growth – Mars Rover Data Rates

- Mars rover data rate values:
  - Sojourner/Pathfinder
    - 9.6 Kbs (basic modem)
  - Spirit and Opportunity/MER
    - 25 Kbs (2 Mbs via MRO and 256 kbs via Mars Odyssey)
  - Curiosity/Mars Science Laboratory (MSL)*
    - 32 Kbs (2 Mbs via MRO and 256 kbs via Mars Odyssey)
- Not so dissimilar from Solar System missions:
  - Smaller antenna sizes and omni-directional transmission
  - Rovers and large (high gain) antennas don't mix
    - Weight, wind, and dust
  - Hence importance of leveraging MRO and Mars Odyssey

* Radio-isotope powered instead of solar

Growth – Total Data Volume

- Estimated Total Mission Data Volumes
  - Sojourner/Pathfinder
    - Less than 1 GB
  - DS1 (Comets)
    - ~2 GB
  - Mars Odyssey
    - 125 GB
  - Mars Global Surveyor
    - 220 GB
  - Cassini
    - 320 GB
  - Mars Reconnaissance Orbiter
    - 20 TB
  - Large Synoptic Survey Telescope (LSST)
    - 30 TB per Night!

Engineering Data – Use

- Engineering data is employed:
  - During development and fabrication
  - During operations (spacecraft telemetry)
- Very specific consumer base:
  - Distribution is not wide, often strictly controlled
  - Time criticality is high in order to support operations
- Data volumes after launch are predictable and a fraction of that of science data
- The challenges with engineering data are:
  - Pre-launch when volumes can go unchecked
  - After the mission when interest wanes, where does it go?

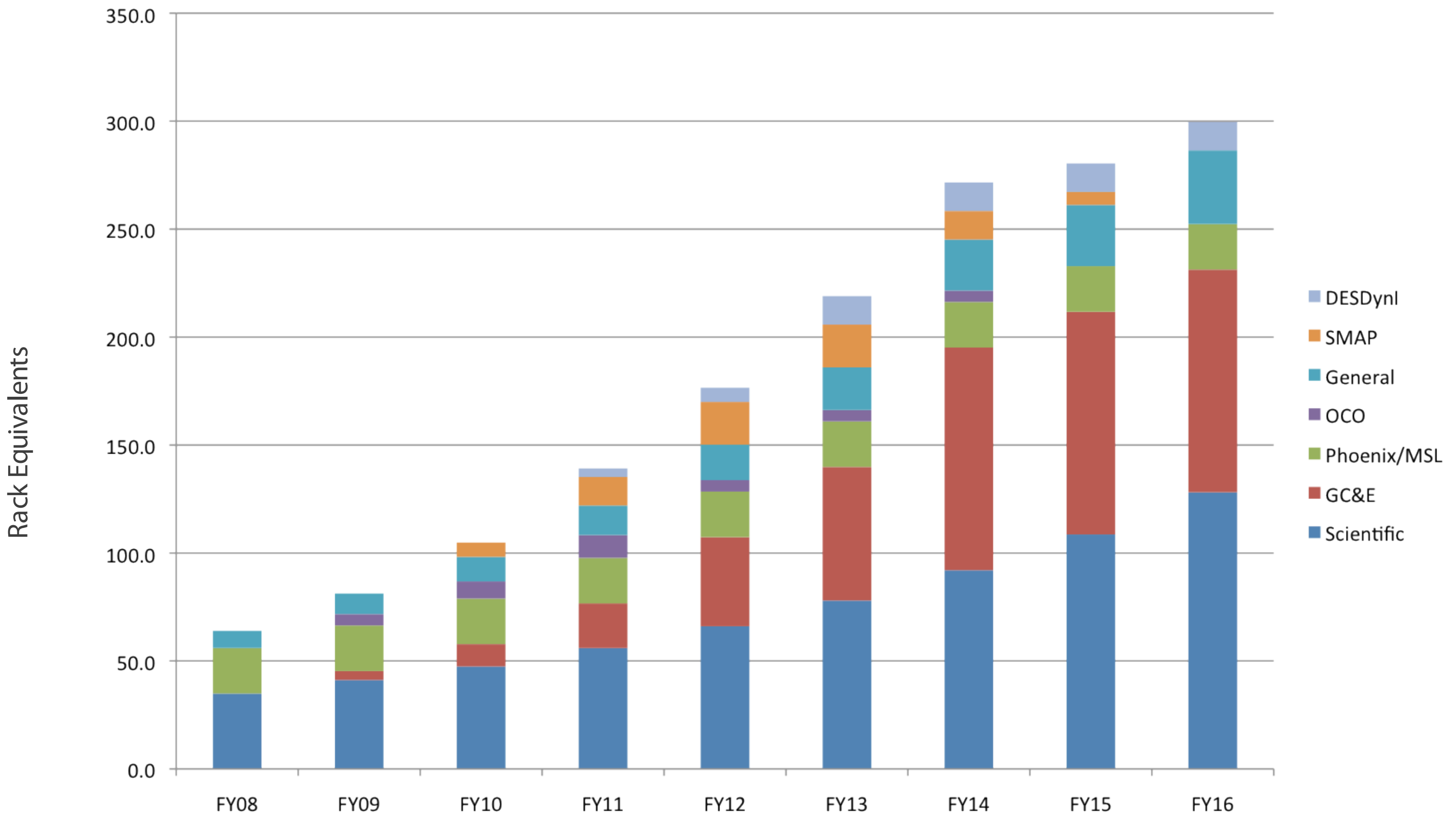
Science Data – Use

- Science data is employed:
  - During the life of the spacecraft and for years after
  - It's why we sent the thing there to begin with
- Also a very specific consumer base:
  - Distribution can be very wide, few controls
  - Time criticality is also high, nothing like the appetite of a scientist who may have waited years for his/her instrument to arrive and fulfill its purpose
- Data volumes are also quite predictable but are heading towards enormous:
  - Extended operations add to the volume
  - Fortunately there are supported archives such as the Planetary Data System (PDS)

Data – Transmission and Distribution

- Both types of data are stored and transmitted by the rover, lander, or spacecraft
  - Source-based volume limitation disappearing with improved technology and leverage
- That data comes to us via the Deep Space Network (DSN) for anything not in Earth orbit
  - Undergoing constant improvement to maintain capability against demand
- It is then processed by the mission's ground data system and then distributed as various products to their consumers
  - Potentially another set of choke points that missions now address

Data - Processing



Data – An Example Future Mission

- Earth-orbiting science mission
- 2-year nominal operations
- 14 orbits per day
- Downlink of 16.3 GB per orbit
- 228.2 GB per day
- 83.3 TB per year
- 166.6 TB for 2 years

Data – Demands on Networking and Storage

- Processing and distribution, driven by data rates and volumes, translate those to pressures on:
  - The mission's ground data system
  - The product distribution chain
  - The storage associated with these
- Network paths and storage system capacities have grown substantially in the past decade:
  - Fortunate because we need that capacity more than ever
  - Some Earth orbiting missions can't transmit all the data they collect during an orbit, instead they have to do on orbit data reduction
  - MSL, in route to Mars, will approach the limits of network and storage performance

Data – A Real Use Case

- Mission "X" required that we design a ground-side network and storage solution unique to their requirements, all centered to support daily planning
  - Understandably they wanted to maximize science collection
  - Bucks the trend of trying to establish common infrastructure
- Data rates on the ground side approaching 6 Gbs were among the specified requirements
- We were able to set up an SLA-driven storage and data transport environment for Mission "X":
  - One that builds on existing shared services infrastructure no less albeit with their own VLAN and dedicated storage arrays
  - Fortunately commodity component subsystems are still viable
    - 10 Gbs Ethernet network
    - High capacity storage arrays with NAS interfaces and common tools (snapshots, deduplication, etc.)

Data – That **Real** Challenge

- However while the physical (performance) aspects are well understood…
  - And addressed to the mission's satisfaction
- The "soft", or people behavior, aspects are not just theory, but now reality, and were not well understood at the beginning of the project:
  - "Where's my data?"
  - "We need large, no larger, No Larger, file systems."
- Here is where "big data" and Proper Prior Planning Prevents Pretty Poor Performance collide
  - So what happened?

Retrieval – Necessary for Use

- You need to get back what was stored in order to use it, at least more than once
- Absent the ability to meta-tag the data the data architecture and storage structure became critical
- However that Proper Prior Planning did not occur, instead previous experience was relied upon
  - For those older, lesser data rates and volumes that might have been okay, even meta-tagged data might have dealt with the issue had it been available
  - However that was not the case
- The data structure lacked granularity and the search tools that could be applied (like “find”) were simply overwhelmed

Retrieval – Necessary for Use

- The result was a breakdown in data management
  - Multiple copies of largely the same data became an issue
  - No longer needed data was left to languish
  - Various methods of structuring the data in the file systems virally popped up
- Significant rework and data migration were required in order to establish some governance over the data
  - Very time consuming
  - Required lots of communication
  - Some data just had to be left “as is”
  - Storage costs skyrocketed necessitating reviews of lower performance tiered storage for lesser accessed data
  - A cold storage tier was added to the list of options

Recovery – Necessary for Use Too

- Then the inevitable happened before we could complete the migration to a better configuration…
- We experienced a major storage array failure centered around NVRAM preventing an automated recovery
  - On their largest volume of course, Uh oh!
- Normally it would have been “routine” except that the customer’s demands for fewer, larger file systems made any kind of recovery other than from a mirror simply impossible
  - This was one 60+ TB file system, too big to mirror at the time
  - Not the kind of file system you pull off of tape
- We did get lucky and managed a good recovery
  - But only after losing a few years of our life span

Lessons Learned

- Lay out a data architecture at the beginning, one that recognizes data volume behavior and storage solution management limitations
- Absent meta-tag capabilities provide for cataloging on ingestion, i.e. have the process delivering the data create some sort of catalog that will facilitate the find and retrieve needs in the future
- Keep file systems to a manageable and recoverable size
- Employ properly designed linked file systems to give the impression of a larger gigantic single one
  - Rarely is a single file going to require anything like that kind of space and if it does redesign your application

Conclusion

- Just because the technology says you can…
  - Huge file systems
- Doesn't mean you should
  - In fact don't
- Big is different
  - "Big Data" must be planned for and managed in light of what it might otherwise cause to happen
- Plan. Plan. Plan.
- We are used to just setting up file systems and then copying data, adding additional file system as necessary
  - This has to change
  - The planning, the data architecture/management, has to become a profession/skill/art