



# **NASA Earth Exchange (NEX):**

## ***Community Engagement in the Cloud through OpenNEX***

Speaker: Petr Votava, NASA Ames/CSUMB

Team: Sangram Ganguly, Andrew Michaelis, Ramakrishna  
Nemani, Petr Votava, Weile Wang

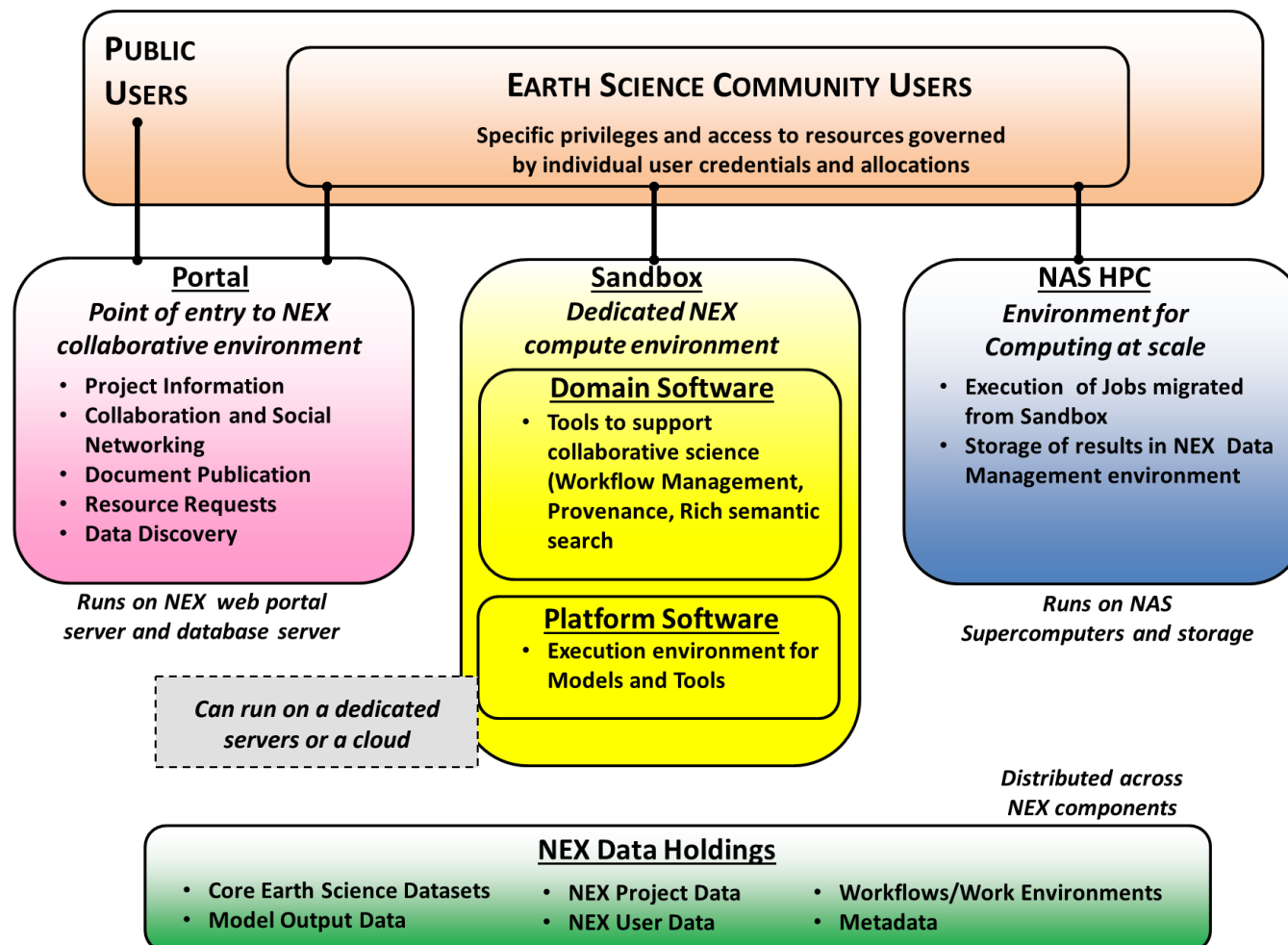
**ESIP Summer Meeting – 7/9/2014**

# Overview of NEX

---

- Supercomputing + data + social networking + knowledge management
  - Part of NASA supercomputing division (NAS)
- Goal #1: Do big things (compute + data intensive)
  - WELD (1.5PB per iteration), GIBS (5.5PB per iteration), DigitalGlobe (PB++), ...
- Goal #2: (Sustainable) Community engagement
  - We can only sustainably drive so much + we are not experts in everything
  - Organize community around data
    - Landsat, MODIS, NAIP, ...
    - Let's see what people can do

# Architecture Overview



# Current Limitations

---

- Resources are not infinite even at NAS and the number of “interested parties” is growing
  - Supporting 30+ teams (some doing pretty big things)
- **Access** and fast turnaround is a problem
  - We want to bring teams in to prototype new algorithms, publish results, provide workflows that others can build on and move on to scale up their work (6-8 months)
    - Collaborative experience – WELD, NAIP landcover, ...
  - Sometimes it takes more than that just to get them on the system
    - Even folks with security clearances had problems

# Expanding to the Cloud

---

- Logical choice – scalability for more teams and not that many security restrictions, no need to build up resources internally just to try things out (risk mitigation)
- Technically not that difficult – we are not providing huge software infrastructure (yet)
  - Move some data, configure some AMIs with tools/libs/data access, scale it up with EMR or StarCluster, ...
  - Lots of info out there – really not anything new
  - We aim just making it easier for users to get started
- BUT: Who will pay for it? For how long? How do we really engage the community? We need to start it before we can try to sustain it
- Goal – prototype in the cloud environment
  - Minimal access restrictions – we only use public data
  - Fast project initialization – we have seen this with HPC NEX
  - Could be fairly cheap for prototype efforts

# The Setup

---

- NASA-Amazon Space Act Agreement
  - An experiment (1 year – till November 2014)
    - Amazon will host the data (public datasets program) and we will engage the science community
    - Lots of Amazon-NASA-Amazon-.. lawyer interaction
      - We can't provide services or any larger pieces of code – everything must be open source and in the public domain
      - This also deals with climate data – a politically charged topic
  - Fast forward 7 months, SAA signed and we can start moving data
    - Landsat GLS, MODIS NDVI/EVI, AVHRR, NEX-DCP30 (downscaled climate) to S3

# Testing and Feedback

---

- Series of contests around the data and services will make it more interesting to users
  - We are trying to engage even users outside the traditional Earth science community.
- NASA Prizes and Challenges Program
  - Start with SpaceApp Challenge (48-hours hackathon)
    - Provided API to the NEX-DCP30 data
    - 22 teams, mixed bag of projects (“Ways to Die” by Drunk-Puppy-Tea-Theatre)
- Improvement
  - Organize around lectures, hands-on labs, usable cloud resources, documentation + contest for \$\$\$

# The Final Product

---

- From past lessons – Supercomputing Workshop 2012
- OpenNEX Virtual Workshop and Challenge 2014
  - <https://nex.nasa.gov/OpenNEX>
  - Lectures (Jim Hansen, Martin Hoerling, + 13 others)
  - Hand-on Labs by our team (for now)
    - AWS basics, AWS + Landsat, AWS + MODIS, ...
    - AMIs setup to go with the labs
    - Wikis for documentation
    - Working on expanding (contribution by others)
  - Two-part Challenge
  - PR (what the heck is a communication plan?)
    - NASA HQ, AWS, OSTP, Innocentive, Twitter, ...



# The OpenNEX Challenge

---

- Run in collaboration with Innocentive
  - <https://www.innocentive.com/ar/challenge/9933584>
- Part one – Ideation (Idea Generation)
  - 7/1 – 7/31
  - \$10K in prizes (at least one \$5K prize)
  - Currently have 220 participants (up from 12 in 2012)
- Part two – Implementation on AWS
  - mid-August through mid-November
  - \$50K in prizes
  - All of you can get involved as well (in both parts)
  - Amazon credits to all registered participants

# What's next - extending SAA and beyond

---

- So that we can do some work ourselves, we are looking to extend the SAA beyond November
- Looking to engage other providers
  - Google (Compute Engine)
  - GoGrid, ...
  - Don't want to be locked into a single provider
- Work with NASA on other ways to procure services beyond SAA
  - Lessons from OpenNEX

# What's next – Elastic capacity

---

- Testing of expanding HPC architecture into the cloud (when needed)
  - Some initial test performed at NAS
  - The key is understanding the tradeoffs between the cost, performance and agility.

# What's next – Visualization + Analytics

---

- Continue on the path of engagement
- Working with Kitware on developing visualization solutions in the cloud that can be easily deployed around OpenNEX data
  - Reusable workflows and provenance management
  - Can be used by others

# What's next - Experiments

---

- Test some of the existing production pipelines
  - WELD, NAIP, ...
  - Look at more details at performance vs. cost
    - Can feed back into our resourcing at NAS
  - What does it take for a large project to be executed in the cloud environment.

# What's next – Easier access to HPC?

---

- Background work for improved compute access at NAS
  - Can span multiple projects (beyond NEX) – really understanding the limitations, many people interested, but really it is mostly about **access**
  - **Extensive work on LOA-2 user verification**
    - Symantec, Verizon, ...
    - Finalizing process and cost
    - If we can't do this, all bets are off for doing this inside
  - Private cloud?

# Some Observations

---

- Technical Challenges
  - There are some, but they could be solved – running in the cloud is not a new concept
  - It may be harder to actually characterize the problems so that you can tailor the architecture
    - Real-time, CPU-intensive, big data, I/O requirements, ...
- Organizational Challenges
  - This is where we have spent a lot of time
    - Where do you even start
    - How do you share your code/workflows through legal means (from NASA)
    - Agreements, security, approvals, ...
- Cultural
  - New ways of doing business – tricky (we are part of NASA HPC)
  - When we started this in 2009 – people thought we were crazy
    - In 2014 NAS is building a private cloud to accommodate some of our requirements

# Ideal (Open)NEX Platform

---

- All projects start with a proof-of-concept and prototype in the cloud
- “Best” projects move forward and are scaled in HPC and/or cloud depending on specific project
  - Seamless to the user = they focus on science and not on the infrastructure and/or data
- Data can be shared between cloud and HPC
  - Community engagement around data
  - Building on others results (reuse)
    - Reproducible publications, reputation metrics, ...
  - Repeat
- Use case - WELD



---

Thank you

<https://nex.nasa.gov>

<https://nex.nasa.gov/OpenNEX>