HPC Thought Leaders Presentations

Table of Contents

search Helpful Links

Questions?

General Support : Miles Kimbrough | miles.kimbrough@nih.gov | 240.276.5251

Consultation and Guidance : Eric Stahlberg | eric.stahlberg@nih.gov | 240.276.6729

Technical Support : George Zaki | george.zaki@nih.gov | 240.276.5171

March 21, 2019

Attendees: Eric Stahlberg, George Zaki, Lynn Borkon, Miles Kimbrough, Randy Johnson, Eckart Bindewald, Jason Levine, Carl McCabe

Notes:

Today's meeting to serve as a working session rather than a share/update meeting.

Looking to begin to capture use cases involving HPC, in a way that we can generate data from it rather than just reports. Capture what people are doing now, qualify/characterize, and start to capture future cases.

Use Case Nice-to-Have's:

Have a **list of accessible use cases** with key stakeholders, to be captured and refined over time; useful for not only HPC but other domains as well (data science). Use this list to iterate with different people in the community, to validate with others in the NIH community, provide a discussion platform of how we are supporting them, and yield better ways to support them in the future.

- Capture use cases and then synthesize to identify CLASSES of use cases
- Ways to identify this community: possible survey (?)
- Program Directors in extramural community are doing some kind of data analysis and trying to get a better understanding of what their grantees are doing. Trying to lift their understanding of grantees' work
- Annotated list – could search for publications which cite Biowulf, text mine publications to search for certain resources

Understanding what's happening in CCR has been a challenge, with 250 investigators using HPC and not always centrally communicating those use cases (pain points, struggles, etc.)

Use cases should also include what people WANT to do with HPC, not just what they *are* doing

HPC Needs Database – what data fields should we start with?

MILES and (if possible) JANELLE – DUE 3/27: Merge Janelle's assessment document with HPC SIG assessment and circulate to team

- ID data fields and what would be expected to be put into responses
- Survey to serve as a tool to be referenced as we go out and engage the community

Earlier discussions around turning those into a database – how would we initially capture info to understand use cases and develop insight. Begin to answer questions around bottlenecks, what's needed – contributors will be able to provide pieces of larger picture

Many scientists have data analysis problem – mostly resort to excel. Need exists to identify communities doing data analysis who aren't equipped with the right tools beyond MS excel.

Provide logistical support to these communities, if they want to host their own seminars and workshops we can provide logistical support (advertising, comms, etc.). CBIIT already has infrastructure in place to help with this.

Concept of *Communities of Practice*

High-level strategy is identifying HPC needs – Goal is to develop strategy for engaging needs as they emerge

Unsure how to allocate funds. Not trivial amount of money and uses, tired of hearing 'this is how it works' – need to find another resource in this case.

Biowulf: Limit on what it can deliver in a finite period of time. 'Large' jobs at NCI/NIH is a 'small' job in DOE terms. Leaves a gap in the 'middle' job size

Propose promoting DOE resources.

Can get list of Biowulf users based by size jobs and start from the top to consult with them for use cases. Also grassroots effort to network organically with users.

Informatics and Data Science Strategic Plan: Potential plan to be developed per Jeff Shilling

(Eckart) Google Collaboratory: useful resource to run python, GPU nodes. Not sure whether able to use or if authorization is needed.

Upcoming Events:

Nvidia coming week of 3/25 for a seminar. Dr. Tang to provide intro and use cases of DL at NIH.

Publicized via list serves, NCI calendar – Randy to forward list serves (Bioinformatics User list serve) - BIOINFORMATICS-USER-FORUM@LIST.NIH.GOV; HPC-SIG@LIST.NIH.GOV; NIH-DATASCIENCE-L@LIST.NIH.GOV;

January 18, 2019

Attendees: Miles Kimbrough, Lynn Borkon, George Zaki, Eric Stahlberg, Janelle Cortner, Eckard Bindewald

Introductions:

Eckardt Bindewald – Sr Computational Scientist, FNLCR, RNA Biology Lab (Sponsor, Bruce Shapiro)

Involved with coding, clustering, parallelization, threading, ML

Writes informatics applications

Janelle Cortner – prior at CCR, now at CBIIT

Big data comes out of HPC

Interest with HPC TL is understanding projections for the needs for more resources, use cases for on-prem vs cloud, making sure we have resources available and planning on the services side

Hoping to discuss Strategic Initiatives related to HPC, want to harness this group as a node for HPC current/future needs

Opening Remarks (Eric)

HPC not just on the computational side but also the data

Amount of data generated in commonplace experiments is dramatically exceeding what we have historically done.

Objective of this group to inform anticipated resource requirements

Need to gather information – both immediate needs/points of opportunity, but also understand how computing is being used – feeding into recommendations

Want to be data-driven moving forward, what we're fostering. The information we gather is dynamic, not static, so needs will continue to evolve as we interact with new/expanding groups

- ID profile of use cases so we can address moving forward

What are the big compute resources available to NCI?

Documentation on available resources is being collected and will be made available

On-prem computing elements:

- Nathan Cole runs large system out of DCEG using in-house compute environment. Systems being run at a division/lab level, but starting to mature and will need replacement/enhancement
    - Competes for available funds for HPC – need to explore including these into broader needs assessment
    - Propose working with Nathan Cole and Jonas Almeida for initial use cases
- System at DCTD that has a large amount of FPGAs being put into it
- MOAB in FNLCR
- Successor in operation now
- Biowulf
- NIH-level and Division-level buy-ins

Cloud computing elements:

- Understanding where data passes through various computing resources. Helps identify/organize use cases by which resources involved with use case. Eric to make this research available
- MS Azure
- Amazon
- Google
- Cloudera
- IBM
- DOE resources – individuals can access. No cost to get started but contains application process. DOE can provide discretionary allocation (10% of available DOE compute)
    - Few orders of magnitude more compute overall

Need more pages geared towards intramural IDing available resources to researchers

Table indicating what resources are available, how large, steps to using, etc.

Data could inform success rates

If need to access human genome data, need to manually download it. May have already been downloaded but not sure where located. Want to reduce redundant efforts

Janelle spoke w Jeff RE: buying persistent storage attached with Biowulf. Biologists for example come in with raw datasets, analyze, and keep data parked in Biowulf – don't need to keep data parked at Biowulf, but some other use cases need to remain parked.

Standard datasets like human genome – lot of researchers need and keep downloading them. Will make for ideal use case to park this information in Biowulf

If can get data in DME, where people are given appropriate permissions, then no need to transfer data. Commonly needed reference datasets would reduce duplicate efforts

When to use MOAB/Biowulf/others?

Use cases to inform

March version of Jack's report contains guidance on this

Need statements have been produced in the past, contain first incarnation of HCP needs but refresh constantly needed as the field moves too fast. How do we capture use cases so they can be shared on ongoing basis?

Even if we had some type of database to capture use cases, what's being done, essential elements we want to know about, and make extensible so if we want to ask new questions in the future, data can accommodate evolving inquiries

Use cases could be what's being done already and what researchers want to do in the future

Not too many heavy HPC users - ~10-15 in Biowulf, not many in cloud, handful using DOE resources, ~10-15 using Moab

- Could capture use cases among these groups since not overwhelming amount of users
- As more and more scientists begin adopting HPC, desire to capture their use cases as well

Another use case – searching human genome by individual. Not possible directly but can download data. Why doesn't NIH provide searchable/google equivalent of human genome sequence?

- From cancer perspective, would be ideal assuming access and permissions for patient level data are met
- DCEG has 300 GWAS cohorts, used internally but not externally shared
- Can we recreate GWAS cohorts, or what does it take to get them shared??
- Significant investment in creating data – how do you incentivize data in a way to make it accessible to others

CRDC (Cancer Research Data Commons) has elements available ie genomic data commons

- Imaging and proteomics commons not yet available but will be
- Goal to be centrally available

Questions to ask about every use case – need to establish a baseline set of questions to understand common needs and trends moving forward

Janelle volunteering to come up with short list of use cases/compute resources on Biowulf/Moab, AND create draft of questions (to be refined)

Eric volunteering to work with BIDS to establish database (similar to Miles/Randy development of TRON, using Filemaker) of use cases

When making recommendations, these are recommendations for procurements – which will result in larger HPCTL attendance/engagement

Want to justify end-of-year procurements because always will be year-end money

Current tools like Galaxy/DNA Nexus can't be used off-premise. Solution needed to bring bioinformatics into every lab

Janelle's experience with DNA Nexus – trained over 200 people on cloud resources centrally funded through CBIIT - Haven't been well-received.

Vishal will present pipeline from DNA Nexus to Pallantir for genomic/RNAseq analysis during Jan 24 Foundations workshop

Part of Janelle's job with Jeff to bring resources available to CBIIT, make case to Tony not to just use 3 cloud resources but also fund DNA Nexus (cloud resource). Might be able to support this.

With Pallantir, users who aren't coders use informatics tools written by experts, can drag/drop/customize, then use APIs to loop out to compute (seven bridges, Biowulf, etc.) then bring back into to Pallantir for analysis

- Janelle to provide demo to Eckardt

Imaging doesn't fit very well with Biowulf, motivation to move to cloud but problems here as well

Use cases will continue to evolve

As database put together, ensure information is presented in a way that captures surrounding environment (HPC needs 'for the rest of us')

Smaller group working on this project to meet in 2 wks, to discuss needs assessments updates

Move broader HPCTL to quarterly basis

If we know what budgets are across labs and experiments to be conducted, if experiments aren't doubling then available data won't double

Get insight on what are planned experiments for the year

Going to individual data users is the best approach

NEXT STEPS:

JANELLE – provide draft list of top 20 users/use cases, and questions to be asked

~~MILES – send Janelle HPC SIG needs assessments~~

ERIC – work with BIDS to justify development of database to cover/organize use cases

MILES – get HPC Utilization report from Jack Collins (Janelle has)

MILES – update HPCTL wiki

JANELLE – Send Eckard link to data commons

Janelle to provide Eckard access to DNA Nexus through Peter Fitzgerald (runs BTEP)

- Contains freebie access

Smaller group working on this project to meet in 2 wks, to discuss needs assessments updates

- Move broader HPCTL to quarterly basis

Attendees: Miles Kimbrough, Eric Stahlberg, George Zaki, Randy Johnson, Jack Collins, Dianna Kelly

Intro (Eric):

FY 19 initiatives as added agenda item

Add HPC and AI needs to agenda, collect feedback in context of accelerating data science

HPC and Cloud

What do we anticipate needs to happen for HPC and cloud? What are we doing? What do we need to prepare for?

George – piloting use case with HiTIF, want to have image visualization host in cloud using containers. Cloud team shared prototypes, looking to host in NCI Cloud 1. Per Jeff, can use cloud resources that can act as APIs within the firewall. Working demo on VM, can launch VM to install software, pull data from archive – demonstrates connectivity to internal firewall.

- Potential use case of using cloud not only for HPC processing but also data visualization based on resources using cloud containers (Dockr). Using Dockr to launch software users are accustomed to.
- Don't have to pay for licenses, allows for scalability
- Investigators can launch one of these instances on demand

Jack: CCBR, NCBR working with various groups to put pipelines in the cloud, request coming from Janelle. Asked Sue Pan for direct interactions and connections to get spin-up from VMs, put down into infrastructure itself as opposed to going through other channels like Seven Bridges. Sue trying to give access to Cloud 1 sometime (date TBD).

Randy/Jack been working on follow-up on HPC usage – extension from Biowulf retrospective report. Beginning to look like a manuscript. Currently contains usage from HPC Moab and Biowulf, storage growth and utilization, and compared costs of doing computation and storing data in the cloud. Contains estimate if using AWS in cloud using 3-year contract with guaranteed provisioning and not turned on.

- ***Goal to complete report by 12/21***

Need to be very careful why going to cloud, demonstrate good reason of what we're trying to accomplish. Cost is sizeable with large HPC jobs

- Are we getting a better service? Are we creating better capabilities? What are the appropriate metrics for comparison? Want to ensure we're comparing common metrics (to avoid 'apple and orange' comparison)
- Per George, Omero Columbus (as one example) requires licensing fee and is not scalable. Omero containers don't work well using Singularity
- If investing more in cloud, can redirect funds from Biowulf to cloud – potential longer-term strategy

Want to determine long-term, 3-year cost. Will require communication across this community to validate costs.

- Question of what engineering do we want to outsource vs. conduct in-house. Will determine operations and maintenance

16-24 cores will be sufficient for a lot of the computation being conducted in near future – relatively modest amount of computation

If we had ability to have back-system along with resources that you spin up. If need to scale, can burst into the cloud and utilize during high-peak computation. Under normal conditions, can contain internally.

- Cost and resource-effective strategic. Can keep confidential information in-house as one value-add

Regarding who pays for cloud, per Dianna: need to ***follow up with Carl as he is in charge of proposal going to NCI leadership on cloud funding***

- Using cloud HPC billed to common acct that everyone can charge to vs. per-directorate/per-group model – supposed to be a common account (CAN)

Based on CPU usage in Frederick – majority CPU storage is persistent rather than temporary

- ~ $100K/month
- If considering Biowulf, just the CCR portion of compute and storage is ~$2MM. Assuming bringing back 10% of anything sent to the cloud

Roughly 6 months since last Biowulf report. Need to maintain 6 month reporting cycle

- These assessments are included in FY 19 needs

As we move toward cloud, will require very skilled **human resources** to help move those applications. Will demand larger group of people to help move and support those applications rather than doing work on nights and weekends.

- Governance plan also a factor to consider

Positioning HPC and Data to accelerate Data Science

Ask the question – how do we want to answer this challenge? In terms of looking at how we position HPC and data to accelerate data science, is that something we want to do in future context of these meetings? Or something we should be looking at more broadly – NCI/NIH/similar organization level?

How do we want to develop a strategy to deliver HPC to accelerate data science?

Cloud is one of the pieces of this consideration – what are ideas on how we can present/prepare a collective document?

ID how the data commons factors into this. Are we putting a lot of resources into data commons, or doing more movement of data into centralized repository?

- Edge computing another trending consideration

Interest in adding topics to this – ***next step to develop short whitepaper to capture some of these topics so we have common vision working towards***

- Aligns with needs with HPC going forward
- Whitepaper could be deliverable under FY 19 plan – people working on these initiatives. Could be summary document, quick read making some of the information broadly available
- Could have some addendums with additional information, presenting just high-level summary with addendums available for reference

Updates Around the Community

George: Do we have enough HPC resources for AI/DL efforts?

Jack working with Jay installing GPUs into nodes to assist with ML. Biowulf has a few GPU nodes but becoming increasingly in-demand.

George recently conducting inference on images. Recently discussed with lab.

Training not currently being done at scale, given available resources on Biowulf

- Not a question of hardware anymore. More about getting/cleaning data, processing, getting into something that makes sense. Once training models, can then go back to PIs to ensure models will work for their research
- Going back and forth multiple times to conduct retraining, cleaning data, retraining data has been larger part of time than putting on the machine. Once on machines, training can be done over a couple days. Other efforts are people-work up front, requires significant domain expertise sometimes requiring data that some labs don't have
- 80-90% of effort is in the up-front coalescing/comingling/aggregating data

George, Yanling, and others met with Nvidia team – several people working with NIH through CRADA, focused on image segmentation on MRI's. Discussed smaller collaborative projects using GPUs, optimizing/benchmarking pipelines. Nvidia capable to doing smaller projects if there is value and not trivial. Looking at how to share data/software with Nvidia using Biowulf or other avenues

As working with individuals with Nvidia who have access to Biowulf, where do we go for guidance about policies of sharing data with that group? Since they aren't employees of NIH – essentially consulting? OPEN question…

- Is it even possible to share data with Nvidia if they have specific access to Biowulf?

Randy: Recently had HPC SIG meeting on 11/27. Mark Jensen gave talk on creating tumor-normal classifier. Next meeting scheduled late Feb/early March… topic pending

George: Coordinating with Tom Misteli to develop online presence NIH.AI on NCIP Hub

- Developing focused min-workshop on image segmentation Feb 14 – roughly 3 hours with 3 speakers, technical overview. Review of what worked, pitfalls, pre and post-processing pipelines, and encourage attendees to share pipelines on GitHub, to provide overview of resources used

Eric: NCI-DOE Governance Review Meeting being held February

Another workshop proposed has been accepted: HPC Applications in Precision Medicine @ International Supercomputing Conference

- Much broader slate of European participation

SuperComputing 19: Workshop proposal deadline in ~6 weeks. Workshop topics currently under review

Education – Data Science Oriented Workshops and Training

Miles: Upcoming Foundations of Cancer Data Science I Workshop to be held end of January. Providing high-level opportunity for those who've heard about data science to get topical overview. Talked with Cancer Data Sciences laboratory

- Whole effort being defined of filling in gaps of mapping other data science training activities going on to develop integrated master schedule

Emily and Carl working on overall data science training and education effort, partly a response of various reports coming out from Big Data working group

- Education/training resources important part, but people and data part are the rate-limiting factor

Randy working on things such as Programmer's Corner and HPC SIG

Need arising to determine where there is knowledge and where there are gaps – ***where do gaps exist and where are outlets to fill those gaps?*** What gaps exist that need to be filled?

- Randy: haven't thought about this in depth. Recent interest expressed in analyzing RNA seq data better, put in context of gene pathways and visualization
- Jack (in ABCS context): Human resources are critical to training researchers as they begin learning about new tools and technologies. Desire to have a place for consolidated training resources.
- Outreach part of letting people know what's possible, and having consolidated place where people can submit request targeted towards those who can offer consultation and guidance. Then need to have the people with broad knowledge and bandwidth

Action Items

Begin capturing prospective needs – what we're trying to accomplish, what goals are, how we plan to achieve them. Options to consider – integrating in-house and cloud capabilities

Develop short whitepaper to capture some of these topics so we have common vision working towards

Meeting frequency: Propose moving to quarterly and extend as needed

December 20, 2018

Attendees: Miles Kimbrough, Eric Stahlberg, George Zaki, Randy Johnson, Jack Collins, Dianna Kelly

Intro (Eric):

- FY 19 initiatives as added agenda item
- Add HPC and AI needs to agenda, collect feedback in context of accelerating data science

HPC and Cloud

- What do we anticipate needs to happen for HPC and cloud? What are we doing? What do we need to prepare for?

- **George** – piloting use case with HiTIF, want to have image visualization host in cloud using containers. Cloud team shared prototypes, looking to host in NCI Cloud 1. Per Jeff, can use cloud resources that can act as APIs within the firewall. Working demo on VM, can launch VM to install software, pull data from archive – demonstrates connectivity to internal firewall.
    - Potential use case of using cloud not only for HPC processing but also data visualization based on resources using cloud containers (Dockr). Using Dockr to launch software users are accustomed to.
    - Don't have to pay for licenses, allows for scalability
    - Investigators can launch one of these instances on demand

- **Jack**: CCBR, NCBR working with various groups to put pipelines in the cloud, request coming from Janelle. Asked Sue Pan for direct interactions and connections to get spin-up from VMs, put down into infrastructure itself as opposed to going through other channels like Seven Bridges. Sue trying to give access to Cloud 1 sometime (date TBD).
- Randy/Jack been working on follow-up on HPC usage – extension from Biowulf retrospective report. Beginning to look like a manuscript. Currently contains usage from HPC Moab and Biowulf, storage growth and utilization, and compared costs of doing computation and storing data in the cloud. Contains estimate if using AWS in cloud using 3-year contract with guaranteed provisioning and not turned on.
    - ***Goal to complete report by 12/21***
    - Need to be very careful why going to cloud, demonstrate good reason of what we're trying to accomplish. Cost is sizeable with large HPC jobs
        - Are we getting a better service? Are we creating better capabilities? What are the appropriate metrics for comparison? Want to ensure we're comparing common metrics (to avoid 'apple and orange' comparison)
        - Per George, Omero Columbus (as one example) requires licensing fee and is not scalable. Omero containers don't work well using Singularity
        - If investing more in cloud, can redirect funds from Biowulf to cloud – potential longer-term strategy
        - Want to determine long-term, 3-year cost. Will require communication across this community to validate costs.
            - Question of what engineering do we want to outsource vs. conduct in-house. Will determine operations and maintenance
            - 16-24 cores will be sufficient for a lot of the computation being conducted in near future – relatively modest amount of computation
            - If we had ability to have back-system along with resources that you spin up. If need to scale, can burst into the cloud and utilize during high-peak computation. Under normal conditions, can contain internally.
                - Cost and resource-effective strategic. Can keep confidential information in-house as one value-add

- Regarding who pays for cloud, per Dianna: need to ***follow up with Carl as he is in charge of proposal going to NCI leadership on cloud funding***
    - Using cloud HPC billed to common acct that everyone can charge to vs. per-directorate /per-group model – supposed to be a common account (CAN)
    - Based on CPU usage in Frederick – majority CPU storage is persistent rather than temporary
        - ~ $100K/month
        - If considering Biowulf, just the CCR portion of compute and storage is ~$2MM. Assuming bringing back 10% of anything sent to the cloud
        - Roughly 6 months since last Biowulf report. Need to maintain 6 month reporting cycle
            - These assessments are included in FY 19 needs

- As we move toward cloud, will require very skilled **human resources** to help move those applications. Will demand larger group of people to help move and support those applications rather than doing work on nights and weekends.
    - Governance plan also a factor to consider

Positioning HPC and Data to accelerate Data Science

Ask the question – how do we want to answer this challenge? In terms of looking at how we position HPC and data to accelerate data science, is that something we want to do in future context of these meetings? Or something we should be looking at more broadly – NCI/NIH/similar organization level?

- How do we want to develop a strategy to deliver HPC to accelerate data science?
- Cloud is one of the pieces of this consideration – what are ideas on how we can present /prepare a collective document?

- ID how the data commons factors into this. Are we putting a lot of resources into data commons, or doing more movement of data into centralized repository?
    - Edge computing another trending consideration
    - Interest in adding topics to this – ***next step to develop short whitepaper to capture some of these topics so we have common vision working towards***
        - Aligns with needs with HPC going forward
        - Whitepaper could be deliverable under FY 19 plan – people working on these initiatives. Could be summary document, quick read making some of the information broadly available
        - Could have some addendums with additional information, presenting just high-level summary with addendums available for reference

Updates Around the Community

- George: Do we have enough HPC resources for AI/DL efforts?
- Jack working with Jay installing GPUs into nodes to assist with ML. Biowulf has a few GPU nodes but becoming increasingly in-demand.
- George recently conducting inference on images. Recently discussed with lab.
- Training not currently being done at scale, given available resources on Biowulf
    - Not a question of hardware anymore. More about getting/cleaning data, processing, getting into something that makes sense. Once training models, can then go back to PIs to ensure models will work for their research
    - Going back and forth multiple times to conduct retraining, cleaning data, retraining data has been larger part of time than putting on the machine. Once on machines, training can be done over a couple days. Other efforts are people-work up front, requires significant domain expertise sometimes requiring data that some labs don't have
    - 80-90% of effort is in the up-front coalescing/comingling/aggregating data
    - George, Yanling, and others met with Nvidia team – several people working with NIH through CRADA, focused on image segmentation on MRI's. Discussed smaller collaborative projects using GPUs, optimizing/benchmarking pipelines. Nvidia capable to doing smaller projects if there is value and not trivial. Looking at how to share data /software with Nvidia using Biowulf or other avenues

- As working with individuals with Nvidia who have access to Biowulf, where do we go for guidance about policies of sharing data with that group? Since they aren't employees of NIH – essentially consulting? OPEN question…
    - Is it even possible to share data with Nvidia if they have specific access to Biowulf?

- Randy: Recently had HPC SIG meeting on 11/27. Mark Jensen gave talk on creating tumor-normal classifier. Next meeting scheduled late Feb/early March… topic pending

- George: Coordinating with Tom Misteli to develop online presence NIH.AI on NCIP Hub
    - Developing focused min-workshop on image segmentation Feb 14 – roughly 3 hours with 3 speakers, technical overview. Review of what worked, pitfalls, pre and post-processing pipelines, and encourage attendees to share pipelines on GitHub, to provide overview of resources used

- Eric: NCI-DOE Governance Review Meeting being held February

- Another workshop proposed has been accepted: HPC Applications in Precision Medicine @ International Supercomputing Conference
    - Much broader slate of European participation
    - SuperComputing 19: Workshop proposal deadline in ~6 weeks. Workshop topics currently under review

Education – Data Science Oriented Workshops and Training

- Miles: Upcoming Foundations of Cancer Data Science I Workshop to be held end of January. Providing high-level opportunity for those who've heard about data science to get topical overview. Talked with Cancer Data Sciences laboratory
    - Whole effort being defined of filling in gaps of mapping other data science training activities going on to develop integrated master schedule
    - Emily and Carl working on overall data science training and education effort, partly a response of various reports coming out from Big Data working group
        - Education/training resources important part, but people and data part are the rate-limiting factor

- Randy working on things such as Programmer's Corner and HPC SIG

- Need arising to determine where there is knowledge and where there are gaps – ***where do gaps exist and where are outlets to fill those gaps?*** What gaps exist that need to be filled?
    - Randy: haven't thought about this in depth. Recent interest expressed in analyzing RNA seq data better, put in context of gene pathways and visualization
    - Jack (in ABCS context): Human resources are critical to training researchers as they begin learning about new tools and technologies. Desire to have a place for consolidated training resources.
    - Outreach part of letting people know what's possible, and having consolidated place where people can submit request targeted towards those who can offer consultation and guidance. Then need to have the people with broad knowledge and bandwidth

Action Items

- **Begin capturing prospective needs –** what we're trying to accomplish, what goals are, how we plan to achieve them. Options to consider – integrating in-house and cloud capabilities
- **Develop short whitepaper** to capture some of these topics so we have common vision working towards
- **Meeting frequency**: Propose moving to quarterly and extend as needed

September 27, 2018

CLICK HERE for meeting notes

June 21, 2018

HPC Thought Leaders Meeting

6/21/18 Meeting Notes

Attendees: Eric Stahlberg, Janelle Cortner, Sean Davis, Lynn Borkon, Miles Kimbrough, Paul Fearn

Agenda: Review of data services efforts/initiatives around the room, Open discussion

Opening remarks

Eric:

- Data services underpinning several upcoming initiatives
- NIH Data Science strategic plan review and feedback
- Janelle joined CBIIT with new role

Janelle:

- Running data management, acting as facilitator/coordinator
    - Learn about pieces, perform gap analysis
    - Working closely with Eric and Sean
- Excited about transparent portal where PI's can see where their data is stored
    - tools exist with Eric's support with limited visibility, want to raise awareness and operationalize these tools
    - allow PI's to see where data is stored, develop data mgmt policies

Review of Data Science Strategic Plan

Sean:

- Think of biomedical data science as a loop starting with hypothesis, perform experiments, generate data, etc.
    - data needs to be explored, shared - creating models that need to be tested and translated back to human health
- Data Science plan focused mostly on data SHARING which is only one component of a necessary data mgmt plan - does not address/close the loop of generating and applying models to patient impact
    - Does not mention developing leadership across NIH institutes
    - NIH open position of Chief Data Strategist - former CDS only received 5 human resources to help, not enough support in previous effort
    - c.f. Regina Barzilay (NLP focus)

Paul (addressing data prep & aggregation to develop models):

- Agree with Sean that Strategic Plan is underrepresented and under-resourced/undervalued
- There are PM leadership folks, data scientists (computational linguistics/biologists), software engineering (delivering products/services)
- Run into problem with intersection of these e.g. find PMs that don't have data science experience, Data Scientists with lack of mgmt experience
- Strategic Plan does not address this passing of the baton

Sean:

- People who do this well don't have a baton handoff, rather one interdisciplinary group with data scientists, thought leaders, domain/security experts, software engineers
- Not meant to work independently - success of ONE project dependent on entire team
- Within NIH, need to have 'people eating their own dogfood' - if you make it you have to use it
- Higher likelihood of success if all are involved
- Strat Plan does mention about cloud computing
- NIH/CIT trying to negotiate 250 PB of cloud storage
    - All 6 top cloud providers interested in competing (IBM, Amazon, Microsoft, Google, Raxbase, Salesforce, Oracle)

RE: using cloud services on current CCR projects

- 2 large data projects - 1st analyzing several thousand RNAseq samples on Amazon S3 bucket, calculating means and standard dev.s, doing summaries of content bias - each sample generates 20-25 files leading to 10s of millions of files (10s of millions of megabytes) - 30-40 T total space - using Apache spark (system for sharing memory/resources in a cluster environment, data abstracted in a way that looks like 1 gigantic dataset)
- project working with HCF group and NASA extracting data matrix - living in pieces on S3, can subset using R brackets, allowing for batch analysis of large datasets - current dataset with 180K RNAseq samples, another with 1.3 million single cell samples of human brain
    - looks like one large dataset on laptop
- Other project working on omics metadata - using apache spark for data munging, loading into elastic search system (SRA, Geo, Biosample are in), all running on Amazon
    - These projects to be turned into publications ~60 studies
- Accessibility of projects - process to make publicly available - deliverable to include individual files, R package to allow access of files, HDF 5 approach, can run spark over S3 to access datamodel of parking data in Amazon ($0.02/gig/month), not implementing specific tools
    - download costs to easily outstrip costs

    - HDF 5 efforts not a download, rather paying costs for server use

RE: Strat Plan - to meet in Cold Spring Harbor and develop a proposed/updated data science plan

- ID what makes successful biomedical data science program (funding, leadership, etc.)

Paul:

- Curious of status for JDACS public website
- IRP representation/input on JDACS messaging
- Governance Review Cmte, NCAB-sanctioned cmte advising FNLCR including outside representatives having background with NCI & DOE, Trans-NCI DOE Collaboration Program Team (open seats availble for CCR, DSITP, ensuring extramural interests are included), & Implementation Steering Cmte
- Idea to include Sean in new effort being developed by Trans-NCI DOE Team bringing DOE scientists together with NCI scientists, focus on extramural, to ID challenges and opportunities to move forward (e.g. pulling data out of clinical reports) - potential topic area to get boots on the ground
- Potentially bring in Regina B. to this effort
- **MILES - GET WITH LYNN ON ABOVE EFFORT**

Sean:

- Converging Computational Science with Basic Science - 20-30 individuals to attend a workshop addressing common problems
- come with a scientific challenge that you want to make progress with - hackathon approach - with output of project plan (idea of extracting clinical data from patients not a project but an idea)
- Depending on the audience, plans don't yet exist because leaning into areas that haven't yet been pushed - valuable to have mix of short/long term visioning
- Trying to ID sparkplugs, session leaders, to think about idea of what should be developed in advance
    - Is it too early to focus on project plans or proposals, or are plans closer than previously imagined?
- From June ITCR meeting, many individuals with desire to leverage DOE resources
- subset of people organizing/helping put on October meeting in Newport (AACR Big data/AI for predictions in cancer) - profile of attendees to be considered
- How to leverage ideas that are currently working
    - Input needed from ISC
- ID Shovel-ready, actionable, accelerate-able efforts - take current projects and make them better /faster
- Current effort showing real-time what's happening with 250mm people across US, showing instances when someone gets script filled (where was it, what was it) as a surveillance tool

May 17, 2018

HPC Thought Leaders Meeting

5.17.18 Meeting Notes

Attendees: Eric Stahlberg, George Zaki, Miles Kimbrough, Paul Fearn, Randy Johnson, Janelle Cortner

Opening Remarks (Eric):

Challenge to figure out how to make sure everyone stays mutually informed

- Lot of moving pieces now, both in HPC and Data Services
- Ex. Semantic Infrastructure workshop

Plans to ID role of HPC, Data

Infrastructure, storage, and implementation updates developing quickly

- Need to ID how to help ensure that community has the support needed

Plans for today’s call to open floor for discussion of top efforts among the room

- Miles to solicit discussion items, priority focus areas from Thought Leaders between TL Meetings, to maximize time and ensure key items are brought forward

Randy:

Next HPC meeting on Tuesday 5/22 in Frederick

- Steve Fellini to discuss Biowulf updates

Janelle:

New to CBIIT role, want to ensure connection between scientific users and champion key needs

- HPC keeps coming up in user conversations

Issues arose from users around how Biowulf is working

- Frederick users having issues getting large jobs in the queue

Currently in the process of learning and absorbing

Interested in learning how Moab and Biowulf fit together

- Recent conversations with Jeff
- Moab has distinct and complimentary functions with Biowulf
- Mixed understanding of Biowulf/Moab functionality

Idea of having large compute next to big data in Frederick would be worth considering

- Andrea supportive of this idea
- Potential use of ATRF data centers for Biowulf

Paul:

Significant progress on NLP group

- Semantic Infrastructure workshop recently held

Need to understand connections between people and systems, connect with the right people

HPC as related to Internet of Things (IoT)

- Dovetails with Trans-NCI DOE Collaboration Program Team
- DCCPS POC Roxanne Jensen focusing on connection between HPC & IoT
- Concept of increasing number of sensor/monitor systems creating large amounts of data
    - Need to ID role of HPC in getting value out of those, and strategy/roadmap to work through it

Would like oversight of upcoming meetings/workshops of interest

- Need to ID which people from which groups would be attending which workshops, develop a plan to have POCs bring back top 3 takeaways from workshops to HPC Thought Leaders

With several initiatives outside of DOE collaborations, need to ensure upcoming events are communicated across all

George:

ID how to translate broader impact of CANDLE and related updates to intramural and broader/extramural community

Eric:

End of May: Mini-workshop of Predictive Model coherence, access, and interoperability

- Primary attendees to consist of those working in DL with user stories to share

HPC Program to have additional support within CBIIT ESIB Branch to scale upward, develop broader impact for HPC & Data Services

- Funding assured for next 2 years, future funding pending needs demonstrated in FY ’19 & ‘20

- Funding to focus on education and increased awareness, addressing the 'people' issue, more than infrastructure
- **MILES**: Work with Nina/Comms team to update HPC public facing website **by June 15**

March 15, 2018

Attendees: Trent Jones, Sean Davis, Paul Fearn, Eric Stahlberg, Randy Johnson, Miles Kimbrough

Agenda:

- Updates around the room
- 2019 Priorities

3/15 Action Items:

- **Eric - talk to Jeff about including Sue Pan for HPC SIG meeting**
- **Miles - add Tony and Jeff to future Thought Leaders Meetings**
- **Miles - RE: 4/10 HPC SIG Meeting, discuss hosting for HPC SIG materials (Confluence /NCIP Hub/GitHub/etc.)**

Next Meeting Agenda – Key priority areas:

1. HPC Services
2. Data Services
3. Education & Training
4. Collaborations

Topic areas to cover **current, 2019, and 2020+ priorities

2019 Prioritization

- Training
- Developing services – data and people (current limiting factors)

3/15 – Updates around the room

Paul

- Tring to relate HPC DOE projects to semantic infrastructure
- ID where helpful to make those connections
- 2 areas from SEER program ID'd
    - Generating metadata in a standard way, map those to integrate with other data sources
    - How to use metadata to use different types of analysis
    - Connecting with Betsy, Juli, Emily to ID where this fits in
    - DOE DL models perform well sometimes out the box, sometimes don't e.g. histology (complex problems)
        - Trying to map to human learning, referencing external sources
        - Possible other area of interaction b/t DL approaches and SI, or some other way to bring in external knowledge
        - Goal by 2019 to build hackathon using CANDLE framework with SI
            - Want to be interagency
            - Use different data sources and same toolkits, even tunneling out to individual secure enclaves
            - In process of IDing participants who can bring in datasets to environment
                - CDC, FDA, VA
- Some use cases at NCI are different within government as opposed to outside
    - Want to leverage innovations outside government
- Possibly looking at ORNL as location
- Trying to expand awareness of CANDLE infrastructure, intra- and extramurally
    - Gaps still exist within community on training and education
    - Application with path and radiology reports
    - Example communities behavioral research program as they link various types of data among diverse databases
    - HPC/Biowulf offered as resource to Paul from Sean
        - George can provide consultation on using Biowulf

Sean

- Approach of thinking about tools is valuable when have non-data driven problems
- When you have data-driven problems, all about data rather than tools
- Recommend focusing on datasets over tools and use cases
- Looking at tech refresh for Biowulf systems
    - 12,000 cores
    - More critical issue = NCI data storage over compute
        - Requires more boots on the ground to implement
        - April Hackathon @ NCBI
            - Data Science SIG involved
            - May Hackathon @ Google – Sean hosting with sv.ai (nonprofit looking at leveraging extra hours from Silicon Valley data scientists)
                - Last hackathon had 400 applicants, 150 accepted and attended
                - Focus on popular Reno cell carcinoma
                - Google supplying hardware/data processing power
                - Emphasis on ML, API driven analytics as a service approaches
                - Conducting ML tutorials – one at Purdue and Coal Spring Harbor
                - Workshop at joint statistical meeting – late summer
                    - Focus on cloud based distributed workflow systems and underlying technology that enables them

Eric

- Are there datasets that can be made publicly, even to instruct those how to use
- Possible to use synthetic data
    - Heavier lift b/c not quite realistic enough, models just learn the generator
    - Possibly set up enclaves where people bring in own datasets with constraints, not sharing publicly but having within same environment
        - Data accessible but not shared
        - Using NCIP Hub to build CANDLE NIH community
            - Currently being developed
            - DOE collaboration Ad Hoc Working Group Meeting
                - Held early March
                - Can't release specifics, interest and supposed support confirmed overall
                - Discussion with Eric, Jeff, Robert S to develop a data strategy for NCI
                    - Expecting developments over next few days to ID path forward
                    - Want to get strategy together to ID policies
                    - CBIIT data services – Sunita Menon new technical POC
                        - Initial discussions occurring around where to focus efforts
                        - Beginning of data services team
                        - Durga also brought in to act as another interface
                        - Request within CBIIT to bring in additional resources
                        - RE: funds set aside for training e.g. cloud storage, hpc/cloud systems, metadata /cloud management – within scope for HPC Program but not fleshed out yet
                        - Next HPC SIG April 10
                            - Recommendation to include Sue Pan or someone from web apps development team, as they are now taking over cloud services
                            - AGENDA ITEM: Discuss where to host HPC SIG materials (confluence vs NCIP Hub, etc.)
                                - Recommendation to use open-source platform like GitHub
                                - Confluence/NCIP Hub not necessarily on people's radars
                                - GitHub forces people to focus on content first b /c no content provided, forces people to focus on code and text first

- GitHub presents no barriers to entry
- HPC needs in Frederick
    - New document developed

Trent

- Been using different python dev environment using anaconda, anaconda not required for CANDLE stack
    - Anaconda has scientific libraries and Jupyter notebooks

Next Meeting: April 19, 2018

February 15, 2018

Attendees: Dianna Kelly, Randy Johnson, Eric Stahlberg, Trent Jones, Paul Fearn, Lynn Borkon, George Zaki, Miles Kimbrough

Review of HPC Long Range Plan

- Under development for 2/22 HPC LRP Presentation
- Biowulf now 4 times larger than it was in 2015
- Object Storage – NCI/CCR/NIH-level now has OS connected to HPC systems
- Improved scalability for HPC DME ('generic data lake', descriptions likely needed to include in LRP
- Globus endpoints at NCI now available that weren't available in 2015
- Suggestion to include cloud resources, cloud services, and commons in LRP slide 13, to align HPC capabilities with Semantic Infrastructure (SI)
- HPC outreach – why is focus intramural? Is there policy restricting outreach beyond IC?
    - CBIIT currently operating/supporting internal infrastructure so main focus is IC
    - Possibly explore breaking down barrier to extend CBIIT outreach more broadly
    - HPC environment currently not being leveraged extramurally

Next Meeting: March 15, 2018

January 18, 2018

Attendees: Eric Stahlberg, Randy Johnson, Miles Kimbrough, Lynn Borkon, Paul Fearn, George Zaki, Jack Collins, Trent Jones, Dianna Kelly

Updates

Eric asked to provide updates to the HPC long range plan in February, presenting to Jeff

- Plans to reach out to folks to approach presentation as a team and potentially push back presentation by 2-4 wks (potential due date 3/1)

Request for more resources approved within CBIIT, delivered to NCI Director Ned S. for consideration

Need is for more expertise and personnel, not necessarily compute

- Balanced from expertise in HPC, data services, and relationship building with other DOCs
- Need recognized to mature HPC program out of the startup model

HPC plan includes support of data services, providing means for individual labs to deposit data, annotate, aggregated within a common interface

- Provides standard medium for all to work through

Mid-December, begun discussing with Biowulf the physical extent Biowulf can be expanded, to leverage computing relative to the data

DCCPS has data sharing working group, have troubles sharing data across regions

- With NCI-wide ability to store/host extramural data across broadly supported commons, and coordinated commons has access to data

- Jeff S. most interested in internal infrastructure, scope of DOE collaborations to be inclusive of all available resources

Several efforts building out commons of different types (ex. GDC), actions to build out additional commons, where data is centralized but individually coordinated

- Been talking to several groups at NIH, not just NCI, to coordinate in support of MoonShot
- Trans-NIH Working Group to help address HPC commons strategy (**ACTION to follow on this – Eric, Paul, Jack**)

Issue of moving data vs. moving compute

NCI DOE Collaboration Updates

Trans-NCI DOE Ad Hoc Working Group established

- Sanctioned by NCAB to advise FNLAC about NCI DOE collaborations
- Will come together soon to provide guidance
- Not limited to scope of what pilots have been but broader, DOE has more to offer than what is being leveraged currently
- Focus to be on accessing computing available for non-pilot efforts
- ID what can be done to support NCI workforce managed by NCI as opposed to workforce being outsourced to DOE

DCCPS Still very underpowered on the Data Science side

- Desire to develop and sustain compute efforts across DOCs to stay tuned into forward vision /forward technologies

Training programs needed for frontier technologies based on massively disruptive nature of DOE capabilities

- Not a one-time training investment but ongoing renewal of workforce
- Provide additional background of support to existing personnel and new personnel alike

Pilots developing and delivering capabilities, question arising of how best to translate and transition those

- Effort underway to determine how to extend impact of outputs across NCI DOCs
- Need to be escalated at the institution level, not just CBIIT, based on nature of disruptive technologies being explored through DOE pilots

Several pilot efforts approaching computational and infrastructure demands but lack of integration across DOCs

Biowulf Update

- Biowulf to have latest GPUs (P100s) in near future

Needs and Issues around the room

Jack

CSSI as coordinating center, ongoing effort to build infrastructure and software

GPUs – Lower-priced GPUs performing better across multiple frameworks than high-priced models

- Particular example including image analysis
- Structural biology/molecular dynamics/floating point operations – P100's working best

CPU – Skylink performance has been great, not many tweaks needed

Frederick been collecting data/info on HPC needs, how many need to stay local vs. move to other locations

- Hope to have enough info for a report within a month

Dianna working with Jeff to come up with HPC strategy for NCI, to extend towards NIH level

- DOE is great learning experience but need exists to develop expertise internally (can't rely on DOE forever)
- DOE focusing on next disruptive computing technology, will then move on to the next, need arises to operationalize what has already been developed
    - DOE not going to be a cancer shop, more focused on nuclear stewardship

        - Strategic need to build support to help NCI justify investments (personnel, data capabilities, compute) such as DOE collaboration
            - People need to come first, top priority

Diverse computing needs, supporting many workloads simultaneously along with doing individual large runs, complex demands

George

DME

- Added updates to Cleversafe – experimented with Biowulf to Cleversafe with speeds of 2.5gbps
- Scaled HiTIF into production
- No further restrictions on Globus applications account, no need to share syncs together or wait for transfers
    - Ready to be scaled to multiple groups (bottlenecks from data/storage/compute still exist)

Paul

Storage

- Desire to connect team with Liz Golanders (sp?) to ID ways to promote data sharing (**ACTION**)

Tutorials on using NLP within CANDLE framework to be developed

- Possible integration into CANDLE workshop
- **ACTION TO MILES**: Provide as much forward notice for upcoming workshops to Paul/DCCPS
    - CANDLE day 1 = how to | CANDLE day 2 = how to with hands-on
    - Data not required, datasets available to gain initial experience
    - 2017 CANDLE workshop was NCI/NIH focused, February workshop to open more broadly to public (by invitation only)
    - Clinical data processing, NLP communities interested
    - **ACTION TO MILES: Plan coordinated workshop with Paul/DCCPS to schedule workshop in July/August timeframe – focused on DL with clinical data/NLP**

From hosting to transmission of data, how to keep it protected

- In workshop setting, do we bring data to compute or do the reverse?

IP Management for scalability

- Ongoing challenge of data ownership
- No way to underwrite liability of risk with data shared since can't ever be fully deidentified, policy needed to be developed

Randy

HPC SIG

- 1.17 Training included using Singularity containers on Biowulf
- Suggestion by Yvonne to have Yanling speak on image processing/DL

CANDLE community – **ACTION** to develop CANDLE-specific user group

October 26, 2017

Attendees = Eric Stahlberg, Trent Jones, Nathan Cole, Randy Johnson, Miles Kimbrough

Agenda = Scaling HPC strategy and engagement across the NCI

Opportunity to reset with recent personnel transitions

People = limiting factor impacting HPC strategy engagement

- Resources needed to train and engage community

- Computational experience still a gap to be filled

FNL employees now able to use Biowulf (NIH main system), transforming HPC capabilities

- No longer limited to using MOAB system in Frederick

New machines coming online to take images of molecules, generating terabytes of data in short periods of time

- Leading to increasing demand in HPC and data storage

In internal clusters, would be better if more available – always desire to scale up but need to find balance of capabilities vs. budget

- Recent need for bigger memory nodes

- ~50 nodes running 256 gigs of memory, another 50 with 512 gigs, still not enough

  - Large scale imputation causing overload
  - 100's of thousands of samples running across a dozen chips, becomes complicated
  - Microbiome = other issue causing overload

- Tools being developed by biologists/scientists but not computer scientists, leads to scaling issues as most solutions developed for limited use

- Going to do a tech refresh on older nodes with upcoming move

  - Plan to bring on 8 3T nodes

- Personnel - CGR fully staffed with bioinformaticists, fair amount of scripting occurring but not much software development

  - Preference to prioritize computer scientist that works on cancer rather than cancer researcher that works on computers

- Data storage – fairly comfortable with current resources and ability to offload to local resources to remain in steady usage state

  - Existing archive ~2 PB, would target 50-100 TB/month à .5 PB/year storage requirements
  - Desire to get some type of universal NIH/NCI archival system in place
  - Versions of this are starting with Cleversafe technology, would be helpful to fold in CGR requirements to understand orders of magnitude of storage to ensure longevity and access
  - Observations increasingly becoming deeper, more numerous, and longitudinal leading to more data requirements

Where to put storage, compute, and what are constraints on the network?

- Most budgets not designed to support types of networks geared towards impending requirements

- Leads to possibly having data centers with compute in close proximity to circumvent limited network capabilities

DOE ECP project – Exascale solutions for microbiome analysis – potential ties to Nathan's group

- Possible interest in utilizing

- Next step to connect Nathan with DOE POC to discuss potential to leverage

NCAB ad hoc working group looking at opportunities between NCI and DOE

- Computation to be one of the first areas of focus

- Goal to build bridges, provide additional insight on focus areas

Public availability is key consideration with leveraging DOE capabilities

Trent – Emerging storage, metadata tagging were big items at 10/19 data summit

- Upcoming solutions have object data store model with files and annotation associated, in technology-agnostic way for increased vendor compatibility

Next Thought Leaders Meeting date pending

September 21, 2017

Attendees = Paul Fearn, George Zaki, Miles Kimbrough, Randy Johnson, Nathan Cole, Jack Collins

Agenda = Round of updates and open discussion

Updates

Paul – Projects with SEER program (NCI DOE collaborations), as datasets are assembled they seem to be hosted in different places (subcontractors, etc)

- Similar efforts with VA (NLP, ML) using similar datasets
- How can we take lessons learned from this to be portable across environments
- Want to create linkages b/t DOE, using 4 different registries, and VA
- Can environments be configured such that a model developed on one can be portable across others

DCCPS has upcoming workshops on data sharing – 10/18 & 11/8

- ID barriers to move data to other environments
- What are requirements for data sharing, to be explored

Q: (Jack) Possible overlap with resources developed at CSSI along with collaboration with Army

Paul – Internal meetings will identify common language, ID data use/sharing requirements, level-set process

Jack – VA shown interest in almost all branches of military

Paul – Hope to ID key issues/datasets by Jan 2018

Jack – FNL, Jeff Shilling and Dianna Kelley inquired to explore workflows on moving data to Biowulf, listing of what experimental instruments need to be housed onsite vs. what can be moved out

- Hoping to state which workflows work well (CPUs, GPUs) over the next few weeks
    - Ex. Electron microscopy

Object Store – Eric/George been looking at DME and Globus, apparent issues with files being dropped

- Although notifications indicate all files have been transferred, apparent loss of ~5% of files
- Long term storage strategy for HPC, solution needed to resolve dropped files
- CCBR pushing lots of files separately, Yongmei's group moving one consolidated 'tar-ball' file

George – CCBR has 200T allocated on Cleversafe, configured for them to use data on API, they just need to put internal process in place to use

- Workflows pending to push data into archive and into processing pipeline

George – CANDLE workflow for hyperparameter exploration, originally used for DL but can be used for black box function

- Currently running on Biowulf
- If user is working on molecular dynamics problem as example, best solution reached by tweaking hyperparameters
- You define function as black box, bash script, define what hyperparameters you'd like to explore and train your system with, resulting in loss value
    - Optimized using bayesian methods
    - Value put into workflow, results then generated and delivered back to user
- All wrapped together in Biowulf
- Currently being used for image segmentations and Pilot 1 efforts
- Built into CANDLE framework

Data Management – Yongmei's group ramping up production use

- Up to 8T data since April, 10K files with ~26 metadata describing each file

- Details in process for how to incorporate these into pipeline

- Globus has limits on how much data can be transferred simultaneously

  - Currently queuing up jobs into manageable chunks

- High Throughput Imaging Facility – imaging instruments generating lots of data, want a system to put into production so data can move from Samba into Cleversafe

  - Need storage allocation to proceed
  - Currently getting scripts to automate registration pipeline

- SBL labs (Yunxing) – desire to use Cleversafe to store data

  - Currently don't have metadata available but aiming to annotate
  - Vault has been allocated on Cleversafe, working with sys admin to start using in prod
  - All hardware in place

Nathan – currently in planning stages, planning to move into building next to Shady Grove

- Everything in ATC storage room (HPC storage) to be colocated in Shady Grove data centers

- Construction to be completed late summer 2019, likely 2020 before relo

Able to piggy back on top of NCI Globus server to get storage mounted and get users started

- Able to use Globus as transfer system between ATC storage (~5 petabytes) to Helix

Randy – HPC User Group planning training on containers

- Scheduling next training session in October

Open Discussion

Paul – Wondering who is doing large-scale simulation and generation of synthetic data

- For SEER program/cancer surveillance, is there alternative to cancer surveillance as a passive process – assembling a big database, alternative to build simulation that builds a synthetic picture of cancer with active sampling to calibrate model over time

Jack – we're generating synthetic data for electron microscopy clusters and testing genomic algorithms – comparing to what we're seeing in intramural research program real data

- Mostly used for benchmarking and algorithm development, not scalable to HPC yet

- Similar to metadata sets which we're doing with Army

Next Steps

Jack to have team write up summary, lessons learned from metadata set for Paul

June 15, 2017

Attendees = Paul Fearn, Miles Kimbrough, Tony Kerlavage, Randy Johnson, Jack Collins, Omar Rawi, Eric Stahlberg, George Zaki

Agenda = update from HPC SIG, Thoughts for FY18, Priorities, Other Items

HPC SIG Update

- Milestones on track, could use more clarity on what specific HPC needs are among community
- How much memory, what kind of connection/processors/more specific resources

- Ability to match needs with applications would be useful moving forward
    - This type of technical info may not be possible for users to clarify based on experience /expertise
- First step of getting people together to build community, Next step to identify common use cases to determine software – understand what software is being run, who's running it, could that be accelerated somehow
- Type of problem/computing amenable to solution, takes research by those who can make translation between algorithms and possible technologies that could add value
- Possible to segment users by technical level to engage further – get understanding of applications, size, etc.
        - Get info about applications by users
        - Find if they're running apps outside environments
- Needs are not just computational but educational, can begin to fill out profile – new folks could focus more on education component
    - Provide online references (e.g. MOAB, Biowulf)
    - Communicate resources across group
- Will have increasingly wide wheel with which to engage people
- Largest proportion of scientific users are using apps that have been developed outside of labs that they've downloaded
    - Limited few that are developing algorithms on their own and making them work in an HPC context
    - New code being written is being written in Jack's group, supporting science folks but not within science groups themselves
- Most software being used is standard off the shelf
- Need to capture needs/requirements at the level scientists can provide
    - Skills and education components need to be developed in sync with needs /requirements
- If enough groundswell, hardware will come from one source or another
    - Trying to get enough people interested in SIG is biggest challenge
    - Will be labor intensive
    - Also need to confirm that everything currently available has been tried and exhausted
- RE: State of HPC needs – been doing every couple years, currently behind schedule and in process of developing updated HPC needs statement
    - **MILES** – **add 2015/2016 HPC needs assessment to public HPC SIG space**
    - Has to be written at the level of the receiver – once we get appreciation of application areas, depending on app areas we build support around the technology
    - Not a one size fits all solution that's fast – recognition that we have delineations (GPUs, CPUs)
- **MILES – send HPC SIG meeting minutes to HPC Thought Leaders**
- Useful to look at previous meeting attendees and where they come from, to explore gaps
- Plenty of contacts not able to join previous meeting, if we develop time/cadence to discuss HPC SIG next steps/education areas/areas of interest – want to bring people together in a way they want to be brought together

Thoughts on FY18

- Budgets starting to come up
- ID areas to be focused on and addressed, what needs to be emphasized/deemphasized/added /cut
- People are the critical piece in whole equation
- IDing staff/resource requirements for FY18
    - Need folks to translate between needs/requirements and technical solutions
- Resources to include staffing/contracting/full-time resources
    - Reflecting needs and priorities of increasing HPC group

Next meeting, July 20

April 20, 2017

Attendees: Miles Kimbrough, Eric Stahlberg, Dianna Kelly, George Zaki, Omar Rawi, Greg Warth, Tony, Dianna

- Goal for next 6 months raising scientific productivity of intramural research program and building scientific computing
- Jack Collins and Eric are working on this goal April – September of this year
- Forming user group coordinating HPC education and outreach to users of HPC in intramural research program
- Helping community gain access to HPC that will accelerate cancer research and raise productivity
- Guidance, insight support- react and have discussion on how we can make this goal successful and maximize impact overall
- Tony – HPC education what's the vision on who will offer the training

- Work with user group to identify – several options include George, Eric, bringing in others that have expertise
- Have myriad of options
- Work with user group to get spectrum and priorities for these pieces
- Computational thinking, algorithmically, putting pieces together is on one extreme down to things that George has expertise in (making applications run fat with graphics processors with leading technologies)
- We have reps along this spectrum, not can do training but can appreciate requirements
- Can fill users in IRP program to clarify their needs
- Tony curious if we have budget to bring external trainer if required
- Might be possible depending on scope and size as we clarify scope of Carl's branch
- Leverage relationships and partnerships with universities and so forth
- Deliverables
- **May** – define initial scope of community for HPC and have a first HPC user group (mid may)
- Getting interactions with users to understand what needs they have
- Bringing in a communications plan (miles will be involved)
- Coherence between Frederick and CBIIT overall
- Getting first list from people
- If you have suggestions or people you believe would benefit from being part of this group – send to George or miles / randy Johnson in Frederick
- Having meeting tomorrow with Randy
- July 2017 – websites updated with more information to support HPC resource access
- And second HPC user group meeting
- Testing if content is helpful to them – or how we can improve it so it is helpful
- Have trusted group of users to give feedback
- **July**
- Second meeting – provide another opportunity for feedback about how communication plan is, elements working, website, etc.
- Set up so we have back and forth in getting input from user community and deliver and get back to them and so on and so forth – iterative process
- **August**
- Nice to have – emphasizing educational resources that are available – investigators would like to learn or be able to learn more about HPC even outside of having a workshop or training class
- User group comes in valuable again
- Can take websites we think might be useful – for those in HPC for many years can fill in many gaps in terms of resources trivial to us but valuable to them
- Or useful to us with gaps to those who have a background
- User group will help validate these things
- Wrap up goal for six months – bringing a lot of information together so we can inform forward for FY18
- Had needs assessments developed over the last two years with HPC initiative
- Good opportunity with user community to update the needs and use it as guidance going forward into FY18 and FY19
- Forecasting as HPC needs and resources to support needs – having access and educated properly to increase overall productivity
- Plan to conduct survey of broader community to get insight into what their awareness is of both HPC and resources, we may find some things we are surprised about availability which others haven't found or located
- Ends up coming back in to formulate recommendations and priorities focusing on user group for FY18

- Questions or comments from anyone?
- Nathan – what level of user are you shooting for? anyone that has desire for experience or has experience or more
- Eric first focus is HPC users right now
- People who are dabbling because they think it will help them or power user aspect?
- Where within SWOT we are shooting for
- If we have people across that spectrum let's get their names in and qualify and anticipate level of use at this particular time – useful insight
- Let's not limit it – user group will focus on those who have demanding problems done faster but will change as we pull community together
- If you have a lists of potential users or users let's get their names and get perspective and what you are able to assess as their potential use

Two general categories

- Users already using HPC and can use additional level up training
- Those who are not using it today (maybe bioinformatics folks around who need education around possibilities in using HPC)
- These are two completely different groups
- Let's get a list and qualify them and figure it out in terms of the spectrum

Steps or dates associated with that? Not specifically yet. Randy will help with PM work on this and we will get schedule for it

These are deliverables and milestones we've identified:

May 15, 2017 – Defined scope of initial HPC community; Have coordinated and held first NCI intramural HPC User Group meeting.

May 31, 2017 – Have scheduled and held first education opportunity involving HPC programming concepts; Have established HPC User Group communication mechanism and initial intramural HPC User Group communication plan.

July 1, 2017 – Websites updated to provide visible points of access to information on utilization of HPC resources available to the NCI intramural community.

July 30, 2017 – Have held second intramural NCI HPC User Group meeting

August 31, 2017 – Websites targeting NCI intramural community updated to provide HPC learning raesources accessible to the intramural community.

September 15, 2017 – Results of intramural community survey on HPC awareness completed and summarized in a report. An updated 2017 intramural HPC Needs Assessment in the form of a report. Recommendations and priorities for FY18 intramural HPC User Group activities in a written summary.

March 16, 2017

Attendees: Miles Kimbrough, Eric Stahlberg, Dianna Kelly, George Zaki, Omar Rawi, Greg Warth

- From Dianna's perspective, sharing things she's heard and things she'd like to see in context of HPC and initiatives now and going forward
- Getting the word out to people to have them learn about HPC capabilities that already exist as well as capabilities that are beyond what people are experiencing.
- Eric: many investigators have perceptions about what HPC was and how unattainable it was and as we started to talk about what was possible their perspectives on the type of science they would do started to change.
- Some challenges identified upfront at Frederick when talking about expansion of HPC- not having software engineers to implement algorithms envisioned. What are the scientific questions that one would pursue and ask?
- They would use HPC much more routinely if it were available – this would require more cycles and more support
- Moving forward, be ready to respond to expanded requests that people have for things they would like to do.
- Dianna –cloud have basics in place in terms of contract people, services and security frameworks by the end of FY17 or by end of calendar year – this would help us plan for what resources we would have available or how they would be taken advantage of.
- George – education – what kind of specific steps should we do to get awareness across NCI?
- Eric –one of things is building a user group of people interested in HPC (something we've been doing), and taking information that we have developed and put online and making sure its directed at the investigators and going out and talking to them. Identify a specific list of core group of HPC users and build a community around that particular group.
- Thought leaders and scientific applications of specific needs and having meetings potentially on a quarterly basis of people who are actually using HPC not just supporting it.
- Dianna – based on notes there would be two different groups, one training group and an interest group. A group that does bioinformatics and would require training and another that just needs resources.
- Eric – a third need is about getting the community together as well. Making sure we get clarification about what's going to be effective and useful and meaningful but not necessarily limited to only these people. Access is for everybody – education, etc. Clarity of priorities would emerge as well. Education and where to obtain access to HPC is a priority
- How to use it effectively, how to access it and who are the colleagues using it you could go to for ideas, questions, answers (not exclusive to those who have direct experience) – these would all be outcomes.
- Greg – What are things we would not overlook as needs or key aspects to address for HPC?

- Helping people set up storage, use it more efficiently, help them categorize it so they can search in an easy way, training people and provide resources to script what they need to do (people who don't have the experience). Programs are being stifled by volume of work and amount of data being generated – inefficient on how storage is used and money is being lost in managing it.

- George – demo for Starfish being set up by Bob Leopard – George will work with him to set up dates for the demos.

- Eric – in general working on an updated plan getting pulled together and laying the foundation. We are in process of updating the needs document and getting all this captured and sending old needs document to everyone in order to update it with your thoughts. Helps us identify key areas (things may have changed that we have not accounted for, so we need to capture that).

- Also getting input as we start looking at different areas coming to light that might have an opportunity or need for larger scale computing. Eric and Jack working on that as they build out the HPC user group. Any insight you may share on that would be helpful.

- George – is there a timeline for HPC user group?
- Eric – we are looking at having it framed out in April and May – pulling people together by early or mid-May. Summer is a good time because investigators have new students and interns available to explore new things to supplement their primary direction or accelerate it. Between now and mid-April we will be collecting information on who we want to target and bring in. In April we will reach out to them and by mid-May have that done.

Updates on Current Efforts

- Miles – one of events most imminent is CANDLE workshop April 18th and 19th – workshop provides overview of candle distributed learning environment. How deep learning architectures for cancer research are being established – cellular, molecular and application – machine models and application for the community.

- We will be promoting this to the NCI and opening it up more broadly depending on space. Will be sending out communications by the end of this week with more information on that.

- Building out communications plan for collaboration with DOE – logistical efforts working with the communications team to establish points of contact and putting a framework together to allow us to communicate more cohesively so there are no conflicting events or meetings or other milestones as we continue to build collaboration out.

- Event – frontiers in oncology 2 – Mid October of this year – planning efforts making sure we have people requesting travel and figuring out logistical items and including the right people at the right times.

- Meeting this year will be in New York – details will be sent to Dianna.

- George – focus on Data Management API user interface – made it easier to talk to investigators and what the basic value proposition is. Giving them a way to try it beyond just doing the work on command line (how API originally implemented). Reached out to three groups (CCR, Biowulf team, Dauoud) – taking list of top data consumers from Greg and getting their feedback on the requirements they are looking for and if its making their job easy. Feedback so far is generally positive and people want to set up test accounts to try it out. Will continue to collect feedback and most common use cases. Biowulf team not interested in using it because Biowulf not meant to be archived but for processing. If investigators have way to archive their data, it would reduce the load on Biowulf.

- Eric – using this as an avenue to get greater insight what is generally required and use cases and things not anticipated. Asked Xen Wu on what type of feedback they are getting so we can share this insight and use for the plans moving forward for the archive space.

- DOE updates – CANDLE workshop coming up, also development of a set of yellow tasks to support NCI DOE collaboration, including pilots collectively known as JDACS and working to support ATOM (effort between NCI, DOE and GSK have defined).

- Intramural – we've covered the goal for next 6 months to pull things together, other things are working through budgetary details for FY17 laid out as we make plans for what we would anticipate in FY18 – being cognizant of implications of the administration budget for FY18

- Dianna – any conversations with Warren or Dwayne this week about pilot yellow tasks funding? NO they have been out of the office

- There is confusion about the color of money around these items. Eric's understanding is that its not from a CBIIT budget but about Warren talking with Doug. This is as much insight as Eric has at this moment. When and how and timing of getting these through is the outstanding question. Somebody needs to say, yes, there is money there to do it.

- Greg – figuring out today we won't be able to fully fund Cleversafe archive so working on proposal to figure that out. Developing something today to get to Eric and Jack soon to figure out what can be done with that.

- Looking across all storage platforms - who are the users and how much do we have – did not count cleversafe and DMF before so will be reaching out to Robert Shirley to understand storage utilization. A lot of people not using storage effectively because of the way it is set up.

- First team meeting tomorrow for data governance working group – 6-month project Greg leading that. Looking at best way to get storage on cloud and a team looking at compliance and a third one on management and services. Greg will provide update as they move forward.

- Figure out how to consolidate and give updates to this group.

- Greg and Eric are the NCI representatives but we will ask Jeff what he needs from his side of the table.

- Try to get updates from the last month of that working group but also figuring out how to loop in Jeff and others.

December 1, 2016

Attendees: Eric Stahlberg, Miles Kimbrough, George Zaki, Greg Warth, Nathan Cole, Sean Davis, Steve Fellini

Agenda:

- Updates from Attendees:
  - Nathan – big data mgmt. perspective and HPC
  - No major change in what we have, DCEG is putting a digital pathology lab, so will have some resources tied up in that, currently demoing and security auditing a piece of software (image repository). Early stages of setup and finding personnel etc. Will take care of digital pathology images that DCEG may have. A large chunk of data, still not scale of NGS
  - Given go ahead to expand server room there – we have lack of space for equipment do the work. Currently working on expanding
  - DCEG research cluster: access to people to run analysis. Need to start looking at whether we are going to do a tech refresh and what the future of this system is going to look like (probably more budget based than time-based, so will start discussions with DCEG leadership to see what direction they want to go in)
  - Looking at sometime before '18… maybe decided in Q1/Q2 timeframe and implementing in Q3 of 2017 (within next six months for action then procurement)
  - Some people did pathological work in DCEG but not in the sense of running that type of lab – additional input from external sources is useful to push forward
- Greg:
  - Doing more things with Cleversafe, though not a lot of usage on it
  - National cryoem usage on it. At least 300 TB to use and would like to put in Shady Grove because no room for storage currently in ATC
- Jack:
  - Working with outside groups on optimizing dynamics codes on GTUs
  - Have opportunity for outside people to come in for training on HPC – need to coordinate expenses
  - Preliminary stages of working with NSF centers and a few with Los Alamos
  - Folks in CGR interest on training with HPC ? they tend to be more well versed so it depends on the users (there will be a number who are interested)
- Steve Fellini not on line – no Biowulf update
  - Generally updates with GPUs that are coming online in January or sometime soon
- George:
  - Candle project: testing benchmarks for release – internally in mid December
  - One of benchmarks is requirement to install different benchmarks on it. Where will node come from for us to try these things?
  - Candle is name of NCI – DOE project for distributed learning environment for cancer
  - How will we operate in this environment and where will resources come from to evaluate this environment and testing benchmarks
  - Deep learning distributed learning environment that will scale and support type of applications that cancer research will need
  - Talk with Greg about it to see what we need in order to get it done
- Eric
  - Update on NCI DOE collaboration
  - Beginning of November there was a review of the collaboration (combined)
  - Will post slides
  - Supercomputing was a week later – showed content of the review
  - Review itself went well – still have open questions in some areas
  - Candle environment – identify opportunities etc.
- Tactical updates on HPC
  - Working on resourcing and support from CBIIT
  - We haven't updated priorities from the time of our plan – we are covered in short term
  - Want to extend to include data services and data support – matter of how we move ahead tactically

- Miles
    - Send report of October and November activities
    - Preparing for Globus Hakathon in January
    - Identified location and working on logistics currently , followed by communications
    - 2 days , accommodating maximum of 100 people – expecting around 80
- Kelly
    - Invidia – January 12th – more information to come soon – deep dive training all day hands on
    - Started scientific software webinar series – twice a month on Tuesdays
    - Keeping it vendor specific for now is perfectly fine – outreach and awareness and having people understand what we're trying to do
- SC16 Conference Discussion
    - Jack: '16 was the year of machine learning and the tools to do that. Brought focus on data science and number of companies eager to work with us. Many interesting application areas, with many people looking to NCI and government for guidance on how to contribute to research – these are opportunities, not so much in technological leaps. How can we apply HPC to biomedical computing specifically?Nathan: echoed comments. Deep learning, showcasing GPU clusters to do deep learning well and algorithms being used, and how current machine algorithms don't expand well into that type of system for efficiency and performance. Keynote focused on cancer research and personalized medicine set tone for where everyone was looking. Dr. Kibbe being on that panel brought a lot of focus on NCI as a leader. Outside groups not typically thought of as bio research groups looking to start collaborations with NCI now to further their field as well as the cancer fieldEric: a lot of interest, had workshop on Sunday with 80 people in consistently. Many people wanting to collaborate in this space and exploring and figuring out how to do so. Precision Medicine has dominant themes of this year, as well as China with the top 2 Super Computers.How do we get ready to take advantage of opportunities?
        - Jack: Most valuable resources we have is data – if we can coordinate data and know where it is and who to share it with (deep learning, machine learning) – being ready to catch it and working with people to analyze it. Having that ready would be a honey pot for many people
        - Nathan: bring to forefront of GDC or something like it – become the go to place for all centralized data would make collaborations much easier even if we are still access controlling it and giving them access and enable them to pull it and use it
        - **People in workshop not aware of GDC existing – put a bigger spotlight on that kind of resource**
- **HPC LRP**
    - **He Between** now and march we will reassess needs and input that as updates to the long range plan
    - **Start thinking of what the needs are in the HPC space and consider whether it is substantive enough to make updates to the LRP – some depends on what we get off the ground in 2017**
    - Objectives and key results – putting support in place for archival and data, exploring new technologies in these areas, etc.
    - Had discussion with Globus – they have an approach working with compute Canada, associating metadata files with original files – can we scope a pilot in that space? Metadata annotation without lifting it in a new archive?
    - Let me know if you have interest – we will be framing this in the coming weeks
- Leading Thoughts into 2019 – and what to lead us through 2018?
    - **Nathan moving end of 2019, assuming it will stretch to 2020 by the time they move**
    - Will have refresh and complete platform
    - Hope is to be able to stand up in new location. Immediate term is a bridge hopefully for next few years
    - Jack: More optimization in bringing tools on board as we start moving more and more to GPUs and whatever else comes to our world. Moving our code to take advantage of this will be a huge challenge. Optimizing code is biggest concern
    - Gateways will be for slow code until we optimize our code
    - Having cross disciplinary development group to help with this is a key need starting in FY18
    - Greg: have to replace HPC storage this year, and next year have to update network on there and it will be interesting. Expect to update to 40G in software defined networking making easier segmentation models
    - Major opportunities and challenges between now and then: keep eyes open.
    - Things we will start seeing in 2019 are hard to imagine but plan for DOE pilots will be 3 years by then, we will expect to have integrated platforms for combined data analysis and new capabilities for simulations at molecular level and atomic level, good distributed learning environment that we can start leveraging that scales across different platforms and maximizing different efforts across the NCI. Hopefully also bigger and stronger workforce in computational and data science space – including algorithm and optimization improvement

Next Meeting – December 15th, 2016

Note – December 15th Meeting Canceled.

Next Meeting – January 19th, 2017

September 15, 2016

Attendees:

Miles Kimbrough, Eric Stahlberg, Kelly Lawhead, Braulio Cabral, Greg Warth, Nathan Cole, Xin Wu, George Zaki

Not in attendance:

Warren Kibbe, Dianna Kelly, Jack Collins

Agenda Review

- Needs and updates around NCI and CIT
- Miles HPC communications update
- Cloud update
- Leading thoughts

Next meeting – October 20th, 2016 (may be rescheduled)

Needs and Updates:

- CCR (Xin Yu)
    - Not much updates or needs at the moment
    - Anticipate in Biowulf storage – trial going on for couple of months, have 50 TB of space there, intention of loading these big files of working space
    - **Note- follow up with Xin Yu about archive needs**
    - CIT (no one there)
    - DSITP (Greg Warth)
        - No updates from Greg's end
        - Couple of different groups trying to do cloud – getting them organized
    - DCEG (Nathan Cole)
        - Finished last of old storage nodes for new ones – increased density
        - Supposed to be on track to move from ACC to new facility towards end of 2019, given we have been waiting for a firm date, we can't make it another three years without taking some kind of action with existing facilities. Tentatively approved to expand server room space to provide additional rack space to be able to allow us to exist for foreseeable future, and if 2019 date would fall through as well
    - DSITP (Eric Stahlberg)
        - Leveraged Greg's efforts to put Cleversafe into play and have an enterprise API with sequencing facility to see if it supports archiving facilities (talk to Xin Yu about that)
        - George attended Exascale training
        - Moving ahead with efforts with DOE collaborations – three pilots moving ahead
        - Priorities for projects and ideas for FY17 – efforts looking at clouds and data services and exploratory computing efforts and predictive modeling (CBIIT leadership still working out what type of things will move forward in FY17)
        - Nobody else has updates

HPC communications update (Miles Kimbrough):

- Setting up operational structure for requests and tracking
- Communications needs and updates
- Walk through of information that is available online
- What is it you want to know and want access to on a regular basis?
- Miles putting more structure from organizational and communications standpoint
- Confluence site (collaborate)
- Objective : to give everyone visibility on HPC program, what we're doing, how its structured, what types of support requests we're working on and ways to get in touch with us and other groups
- We would like your feedback
- Suggestions for improvement
- Most important question: what is the information you want to see that you don't already see on the website?
- Miles walks through the different components of the collaborate site
- Eric: we are trying to align what we're doing within HPC program with what CBIIT is doing with its activities and programs (mapping into that context structurally)
- Page focuses more on project and program activities
- For HPC services, there is a similar dashboard that is under development to track service enquiries and requests
- Addressing needs as timely as possible is an objective
- Another key component of our program and growth is communications
- Putting definition behind communications channels to ensure we are targeting and reaching the appropriate audience as efficiently as possible

- We are still in the process of iterating the various communications channels to make sure there are no stones left unturned in terms of the people we need to connect with who have the resources we will need to effectively grow
- Internal operating model is the working model we are using for the structure of the program – it is in iteration mode. Developing workflow to determine appropriate next step once we receive requests and information from various sources
- Having common language that is available to help define those workflows
- HPC Thought leaders meeting notes archive is on the HPC wiki page
- Various types of resources we are still building from an organizational standpoint: eg. Training, education, expertise, computing, etc. – putting more definitions and structure so we have a single repository of what is available and ways that we are able to help the workflow to handle support request, compute, data, education requests, etc.
- What would be useful to include in a monthly update?
- Let us know what type of information would be useful (tracking particular sites, resources, etc.). we want to make sure HPC becomes a resource to help you with your needs and others' needs

Cloud Discussion

- Take several groups pursuing cloud and bringing them together
- Many use cases with many potential solutions
- discussed possible options for accessing
- Many potential solutions
- Established need to identify cloud requests
- Consensus – need to get a handle on all the situations where cloud is being requested
- Get a handle now before it mushrooms too far
- Defining next steps
- DCEG is doing some cloud efforts (Nathan) – how many? Types?
    - Nathan - There's very few known projects taking advantage of cloud – one of those is data download because that's where the data is housed
    - To be able to download data from amazon instance
    - There was a relatively large number of whole genome samples sequenced that just finished – in process of receiving them now and those are going to get sent to a variety of different platforms including seven bridges cloud in order to evaluate best platform to do whole analysis. Outside of that there are individual investigators that have thoughts but nothing heard of explicitly
    - Xin Yu – we have been attending workshops (seven bridge) and another one – at this stage just exploring stage and don't have any real working pipeline using cloud but just learning about pros and cons and what can be done
    - Greg – worries that there are many initiatives going on but no cross-talk. Jeff is writing a white paper to move this forward. We need to be on the same page
    - Latest NIH record mentions NHLBI moving to the cloud – there are clearly solutions out there for doing these things
    - Please share thoughts on defining next steps now or send an email so we can reach out and have a discussion on what we are pursuing for our needs being met by the cloud

Leading Thoughts (opening up for discussion)

- Driving questions
- Driving challenges
- New opportunities
- Emerging needs
- New technologies / capabilities / avenues of research

- Eric – this is coming out of meeting in July – frontiers of predictive oncology – miles coordinating developing white paper form that meeting. Context – as we approach cancer from a predictive perspective and see what is happening in computing what do we see as opportunities and challenges going forward?
    - Patient leads – what we do has to ultimately impact a patient in some context to make them connected
    - Data is a huge challenge in many contexts – volume and limitations – from too much data to handle to not enough for discussion
    - Need to bring and integrate data together become strong. Some limitations is data are not consistently annotated or well annotated (will require a lift)
    - With DOE – discussions about what does future data ecosystem need to be? Where data is more readily shared? What are sharing requirements and data requirements?
    - Strong push to make data in preliminary context – not pre analysis but pre -publication – push to get team science going and support in that direction
    - Sharing similar experiences with everyone else if you have them
    - Miles – what stuck out from meeting is importance of collaborating with other groups, which will be discussed in the white paper. People trying to solve similar problems. If we have ability to identify similar problems and avoid duplicating efforts and coming up with solutions faster
- Greg – we have a lot of work to do to be ready for a team science environment
    - Downloaded 2.5M files in August and having hard time getting that data
    - Not enough people to support that part of infrastructure and need to do more in cloud
    - Nathan – dropped off
    - Concluding meeting ten minutes early

- Kelly – deep learning class?
    - Michelle Don and ads office sent out minutes in an email – there is a need for another class. Overview (1.5 hrs) and full day session - not sure if full day is hands on (allowed 100 people in class)
    - Trying to figure out if there's a need for NCI to do our own deep learning session
    - If you have interest in deep learning let us know – reach out to Kelly
    - Deep learning is critical way to derive large amounts of value from data

Take this back to your communities and ask and if there is interest share it with Kelly

August 25, 2016

Tentative Agenda

- Updates from around the room including expected efforts in FY 17, and emerging needs – 15 minutes
- Data services updates/discussion – 10 minutes
- Review anticipated priorities going forward – 5 minutes
- Argonne Training updates from George Z - 5 minutes
- Upcoming events and activities of note – 10 minutes
    - Workshops, training, education, hackathons, etc.
- Discussion: Ideas for future efforts/priorities in HPC– 15 minutes
    - Ideas for FY18

Attendees: Eric Stahlberg, George Zaki, Dianna Kelley, Greg Warth, Jack Collins, Sean Davis

Updates & Upcoming Events

- **Diana** - No needs/updates at this point
- **Eric** - Updated budget process going forward for CBIIT
    - No indication of what would be funded and at what levels
    - Should get confirmation within next 2-3 weeks
    - Special projects, initiatives, scope tasks would be performed independently outside of yellow task
    - Miles K and George Z assisting to create an HPC programmatic dashboard with monthly reporting
        - To speak more broadly than high-level task overview

1. a. **FPOC 2017** - Targeting Q2 2017
    i. Early planning phases, preparation in place for travel/logistical arrangements
   b. **SC16 Conference** - November '16
    i. Call for Papers deadline has been extended for Computational Approaches to Cancer Workshop

- **Greg** - Cleversafe update
    - Met morning of 8/25 to discuss timeline updates of Cleversafe (data file storage) and IRODS (metadata)
        - IRODs not completed yet, decision made to not pursue this interface
        - Still need to focus efforts on Cleversafe in the near-term
        - Discussed desire to join Cleversafe/IRODs consortia
- **George** - Want to see if Greg can put us in touch with other groups for testing on the above
- **Greg** - Sequencing Facility Group involved, will include other 4 groups to identify top use cases
    - Greg to schedule follow-up with other groups to coordinate with HPC team and review additional cases

- Meeting scheduled in September to discuss steps involved with Bob Grossman's team
    - Recent power failure which led to failure of several nodes, unable to recover at this point
    - Discussion to move away from object storage for this reason
- **Jack** - Coordinating with supercomputing SMEs for collaborative opportunities

Data Services & Anticipated Priorities

Globus

-
    - HPC team meeting with Jeff Shilling and team on 8/26 to walk through Globus service, walk through process, and discuss updates
    - Confirm lifecycle of data as it moves onto Globus environment
    - Miles to schedule follow-up meeting in response to latest data request from Parthav

        - Following 8/26 meeting with Jeff's team

- **Jack** - Need to find more efficient method of transferring data rather than bringing in hard drives and dumping into central repository
    - Could be Amazon cloud but process workflow will end up being vendor-specific

- **Greg** - Sometimes supporting data requests is difficult because of staffing shortfalls
    - Will address resource allocation to accommodate increasing demand for data transfer /storage
    - Need to provide quick service and adequate training

- **Jack** - Educational opportunities becoming available between private and public sectors (NSF)
    - Available online as well as in person - no charge
    - Idea to raise visibility of these education opportunities
    - Jack to identify central website with more information, to provide to George Z
    - Need exists to provide a central repository for educational support opportunity
        - Could be an added element of HPC program

Argonne Training Updates from George Z

- 2 weeks, 65 participants, 25 lecturers
- Hands-on exercises
- Topics Covered:

  –Architectures(Xeon Phi, Pascal)
  –Programming models(MPI, OpenMP, PGAS, CUDA)
  –FastMath (PETSc, HYPRE, Zoltan)
  –Software engineering
  –Profiling and debugging(HPCToolkit, Alliena DDT, Vampir)
  –Visualization(Paraview, Visit)
  –I/O (HDF5, MPI)

- Takeaways: We have permission to share programming models explored during training
    - Collaborating with CIT for workshops to discuss
    - Part of Exascale Supercomputing Project
- George providing additional general opportunities to provide educational support for HPC within the next 9 months

Ideas for Future Efforts

- **Eric** - difficult to concretely plan for 2018 with so many current variables within 2017
- **Jack** - CCR will start recruiting for cancer laboratory, will have expanding needs as a response
    - Will be difficult to predict these needs at this point

- Natural language processing - increased demand for storage
    - Will need to start mining and scaling metadata in order to keep up with scale
- Mechanism is necessary to interact with commercial companies more efficiently
    - Will be instrumental going beyond the September time frame
- **Dianna** - Collaborate with Carl to put together a conceptual design for above interaction model

- **Jack** - Need to have small group of people who can devote time and resources to exploring new technologies
    - Need to have a low barrier to entry
- **Eric** - ATRF, FNL are both good candidates from a physical venue perspective
- **Jack** - Need to start reaching out more to user community, meet regularly so we are up to date on current pulse
    - Once a quarter or every 6 months
    - Probe PI community for anticipated storage demands

-Challenge will be that sequencers will no longer be large-cap expenses

-Need to talk with lab chiefs to best understand future data needs

- **Sean** -Science is a couple years ahead of where CBIIT needs to be
    - Need to make it possible for people to use cloud
    - Something that allows PI to make purchases from cloud
        - cancer clouds not usable by most PIs because they can't buy more cycles upon exhausting their allotted compute time
- **Greg** - Some potential solutions out there, AWS is a potential, could put above on a sub-project and charge back to the lab
    - Should make this model development a priority to accommodate increased cloud storage demands

    - Possibly include Tony and Ishvar (sp?) to provide representation and a voice to move this forward
    - Eric to take this as an action to make sure it moves forward by 9/15 - will involve Dwayne and Tony to put model together

Next Meeting: September 15, 2016

July 21, 2016

Tentative Agenda

- New faces and introductions
- Needs and Updates Around NCI and CIT
- Frontiers of Predictive Oncology and Computing Meeting Updates
- Review FY17 Candidate Projects
- We were on hiatus for a little while
- Important to have these meetings more regularly and keep each other updated and aware of what is going on
- We have new faces and important to share updates
- Logistics updates and coordination support
- Suggestions on other priorities to pursue

Introductions

- Anastasia
    - Important we define what we mean by HPC and big data and what we are aiming for
    - Lots of important things going on
- Miles Kimbrough
- Nathan Cole
- Carl McCabe
- George Zaki
- Warren Kibbe
- Greg Warth (Phone)

Needs and Updates

- CCR: Sean out of vacation: not much insight but one item is looking for ways for longer than 1-year retention for files (Xinyu)
    - This is prime for where Helix is going with Cleversafe
    - Email went out looking for beta testers
    - NCI/NIH data archive policy? Doesn't exist
- Storage for new instruments
    - Need life cycle management system for file retention
    - Not all data is equally valuable after one year
    - Needs: to have a data management instrument in place. Storage needs secondary to acquisition of instrument
- CIT – Nobody on phone to give details
    - Object storage – looking for beta testers (cleversafe)
        - A place to put data that is assured over long term but not necessarily recalled forever
        - For it to work well its better as static data versus data that changes a lot

    - Cleversafe reduces need for backup infrastructure
    - Currently have 1.2 Peta bytes for NCI, around 2 Peta bytes for CIT
    - Initial plan was to use an avere front end for cleversafe – couldn't do provisioning on it. Some speculation over what they will do on it now
    - Cleversafe coming out with their own NFS presentation
    - Storage back end for GDC – we want them to be able to scale considerably. Cleversafe too expensive for this kind of storage. SEF is not ready and having trouble going past certain amount of storage.
    - ***Get Bob's technical team to give a presentation***
    - **Data retention policy: at some point if we store it and they want to access it beyond a certain point they should pay for it (we need to discuss this and plan it)**
  - Biowulf

- DSITP

  - Cleversafe object storage (Greg)
    - Have brought up storage here and its been deployed. One node at CIT, one at Shady grove and one at Fort Dirdy. Going live (guarantee that data will be available there and not destroyed on it) August 1 pending results of final testing. Hooked in regular part of network. Segment set up for us.
    - There is latency but not large decrease in speed
    - On 40Gig back bone, and have 10 Gig available.
    - 2 Peta Byes usable storage
    - If cleversafe does have the NFS mounts to implement in first quarter of 2017 calendar year, working with cannonball to do backups and go to synthetic pulls instead of weekly pull backups.
    - Usually 60% - allowing 2 failures (default) ratio raw to available data
    - Bob Grossman very conservative (we couldn't guarantee him any kind of back up)
    - ***We could get cost down (more reasonable) than for him to be that conservative on how cleversafe is set up. Speak with Bob. (Greg and Eric to get a meeting with Bob) – he has 4.7 Peta bytes and 5.5 available space – at least get to his technical team, not necessarily Bob himself. Send a note to Alisson Heath aheath @uchicago.edu***

  - SC16 Computational Approaches for Cancer Workshop
    - There is a workshop in November – we've put a call for white papers for that.

- CBIIT

  - Archive API – works on cleversafe and other pieces. Ready to talk to S3 in general or cloud, etc.

- DCEG- (Nathan)

  - In the middle of installing another Peta Byte of storage. Somewhat of a tech refresh of oldest original 108 L nodes. Density on them is drastically poor compared to anything modern, and going out of support next year. Sticking to Isolon as vendor. 2019 ACT supposed to move in this timeframe.
  - We need to make sure one of considerations is a data line (Carl)
  - If we are that close, as long as we can get direct lines between new building and shady grove we can codirect equipment
  - HD400s for Isolon replacements. Now have mixture of odor and L nodes. X410s and HD400s for pure capacity.

- Other DOCS – No other updates

<u>Logistics updates</u>

- Communications plan being put in place by Miles

- Ramp up in August and run in September

- Open collaborate page to all members of NCI

- Yellow task that Eric is on ends in September – plan for how to continue this being worked on. Summary report on what we were able to do in first two years

- Overall perspective – Braulio yellow task is where we move programmatic support to

Frontiers of predictive oncology meeting

- Well attended nearly 100 individuals each day

- One room, enthusiasm, good networking time

- Limited range to roam encouraged people to have discussions

- Good insight shared in breakout sessions

- Planning a white paper by end of August to pool all input

- Survey in development – intel was asking how meeting went (being iterated on now) – keep paper work reduction act in mind and get intel to do this

- Planning next meeting – get information out earlier and better

- Blog post – makes sense to do with DOE

FY17 Candidate Efforts – HPC and Exploratory

- Data Services Environment

  - Archive and metadata services
  - Explore integration with GDC
  - Transparency on storage utilization
  - Expanded storage on transfer/intermediate Globus services
  - ***(add managing data retention policy and life cycle)***

- **HPC Support Core**

  - Deepen level of support and education for applications of HPC (connecting the HPC with the science and making improvements there)
  - Front-end development for HPC backend
  - Continue support for compute and data intensive applications engineering and optimization
  - Extend level of HPC resources available (cloud, elsewhere)
  - ***Useful to have future visioning to see how we will look in one year***

- **Cloud Resources**

  - Dev and compute in cloud, data storage, archive, development, etc.
    - **Should talk to NCBI (they are making a push to move all their services into a cloud environment)**

-

- Predictive Models Explorations and Assessment

  - What are implications from computational, data and science perspective
  - There is a big misunderstanding inside NCI about what a predictive model is

efforts are not distinct. They need to be coordinated and aligned overall

*** Describe more about what the purpose of HPC is and less about the infrastructure and the "means to an end". This is part of future visioning, underpinning of "why" we are doing this and what the purpose and impact is.***

April 21, 2016

- Data call: they are specifically looking for HPC cost in server and mainframe category

- Capital acquisitions

- Pull from DCEG in terms of what they are anticipating (Nathan)/ or give a view and we can reconcile it with DCEG

- Eric to give presentation

- Data Call: request for info on HPC – clarification from Karen
- Cost for mainframe and server projections for HPC investments – Sean sharing amount he gets each year to make decisions on

- We don't want to double count things. Biowulf and CIT should be taken into consideration

- Nathan waiting on numbers from DCEG
- Sense of total expenditure from DCEG for last year – we don't have that because it's a new category coming in. No data on HPC

- We don't have a real baseline for new categorization

- Include server and mainframe investment cost to support personnel etc.

- Don't know profile of what DCEG is but we're doing HPC support for them. Don't know size of their plant beyond a few stations in their lab

- Asked for Bioinformatics numbers by Melinda Hicks

- Not clear line between doing bioinformatics vs infrastructure necessary to make that happen- need to draw line between those two to not count dollars that support science

- Important to put cost of infrastructure and people to maintain it – this becomes an IT cost but people taking codes and rewriting them and running them on HPC is scientific computing and we don't want to put that in

- We want to be careful and consistent on how we report on bioinformatics activities

- Eric: trying to break it out into categories with Jeff and Tony – total bioinformatics investment is fine but helpful to define physical plant and what it takes to operate that

- In terms of physicality (cost for power and cooling) no idea what those numbers are (not part of the report). But how much spent on storage probably is something we want to report

- Server procurements, storage procurements, etc.

- HPC costs in cloud getting pulled from Tony – cloud HPC resources

- Having info available and working with Karen to best sum it up

- Making bioinformatics info possible is good first line to capture and to know total bioinformatics cost is good but not from FITARA stance (don't want to report that)

- **Presentation**

  - HPC support for FY17
  - Exploratory Computing – foreshadowed in LRP
  - Data services efforts

Give relative priority to these efforts and identify stakeholders

What amount of resources might be appropriate and when for these resources

What makes sense for FY17? What might we defer for later?

HPC support for FY17

- Eric goes through slide

- Prioritizing activities for next year to know how to allocate resources

- First three are program development

  - Getting NCI involved within HPC community and looking at frontiers of exascale computing – help shape what happens in that area and have interests represented and not just responding to what happens
  - What kind of science being supported through this and give some examples?
  - Eg. DOE pilots – what at NCI needs same kind of infrastructure?
  - Storage and kind of compute nodes more important than number of compute nodes for biowulf – laying out drivers is important
  - Exploring how we make HPC more accessible to the scientists – lessons learned from NSF
  - More tactical: evaluating HPC in the cloud – resourcing, how to do it within context of NCI and investigators
  - NCIP cloud computing HPC integration – how would we actually extend the backend computing support for things in the NCIP cloud
  - **Integrating cloud with GDC (probably not existing cloud pilots that would be here) – GDC is the long term investment. Move more towards that and attach HPC to it.**
  - Two more component to flush out – how to build same kind of infrastructure that integrates imaging and what uses there are for HPC in that (mining and indexing) – secondly, not sure you can use single approach when you move from pathology to MRI to ultrasounds, future extraction

looks different. Another part of it is pulling in more data from HRs and what it looks like and what kinds of HPC are relevant for HR data
- GDC – most is on genomic data
- Other types of data moving into GDC (may or may not directly) but will move in to GDC like environment. May not be able to scale GDC like that
- ERIC: a lot of these are going to be intramurally focused
- Using accelerators to move things through sequencers faster – can you drop card in sequencer and improve their QC throughput (so they don't have to move it to a general computing platform)
- How to use cloud for storing data as a deep archive beyond premises (not in use)

- Cloud storage pilot – what it would take to use the cloud for deep data archive

- Leveraging investment in data service environment

- What happens when you have all meta data and feature extraction – investing at least capacity to do high performance analytics and getting systems that have that capability accessible within context of researchers we are working with (supporting work in that space)

- Cloud container environment – containers operated internally and externally and portability between them

- Micro grants – consider as a way to bring external resources into the effort on small projects and supporting intramural investigators in that way

- Prioritizing these: not much time right now / add other ideas / weigh in on what priorities are

- Maturity of being able to do it (having partners ready to do it) rather than priorities setting. Are we ready to do it? We have plans in place to do these things. Identifying future timeline of when these things would be ready to start.

- Not a matter of ranking them in order

- This needs assessment was built off 2015 and various meetings we've had

- Developing involvement within community to help NCI participate in consortia

- Response to NCI DOE pilots and looking at NCI exascale cancer working group – pull people together more broadly – what are the applications of exascale? Extend beyond that group and get broader input, looking long term and being frank about level of computing investment we need to make. Who are partners for that? What is need? What is demand?

- Training and outreach – supporting education and development of awareness about what computing and data science can do for cancer research

- Need to do more HPC training, developing more applications, helping with those who have large data service needs and extending support on using cloud more effectively

- Need people to help investigators use cloud more effectively (2-3 individuals to cover that space)

- How do we take service now implementation providing request support for HPC support and develop that further so it's a better interaction for the individuals who need that support

- Bioinformatics core interface – did users find it reasonably effective? Yes

- Investigators have an idea and communication was done well, so was coordination

- Last two looking at making sure we have project management support in HPC space as we have more request (becoming more project oriented as opposed to task oriented)

- TPM would have more ability to do technical support and know more about problem space rather than a PM. Needs to be depth of awareness about that (it is negotiable) – potentially someone with a lot of HPC knowledge and can translate to technical team

- For FY17 – taking what we've looked at to develop a basic service API that's at an enterprise level and building it out to have stronger and deeper services. Developing façade on top of what many of our object store technologies are that we might use. Rationale is to provide flexibility that we don't become vendor locked. As things become more capable and standardized, façade will get narrower and narrower and potentially disappear.

- Helpful to lay out some of initial projects and right size whole activity so not get carried away building things without having accurate picture

- Useful to call out number of FTEs required to build things out from a budgetary perspective

- Supporting data service environment

- Dedicated system administrative support for it

- How to look out to extend storage to different types of storage places like cloud, etc.

- Next steps: how best to get input and refine this with CBIIT budget process to prioritize. Give opinions on what we should do, shouldn't do and defer or things not on the list to think about.

March 17, 2016

Attendees: Steve, Nathan, Greg, Xinyu, Dianna, Sean, Carl, Kelly, Eric, George, Omar

Agenda

- HPC Long Range Plan
- HPC Needs Analysis
- Storage and Data Services
- Other Items

- Carl: several people have started using Slack to communicate, maybe we can try it out to keep dialogue going outside of here (it's a collaboration tool)

- Eric: HPC long range plan:

- Contribution from many people
- Finalize* version sent to HPC folks. Greg and Dianna going through the various plans
- Get input – taking the long range plan and making it visible
- Website getting up to share information on activities – makes sense to put the LRP up there as a reference document
- Dianna: working on customer versions to distribute to SMW – condensed version of LRP for them
- Making sure LRP is out and not just sitting there – as a reference tool
- Get input on when to target a refresh on the LRP
- Carl: want to do it annually to avoid it getting stale
- Greg: first quarter of calendar year to align the budget process
- Next January: concerted effort to make updates prior to end of March
- Collect information and collate and distill
- Storage and data management needs with main LRP – is this scope we want to maintain in future or have different approach? Keep this in mind. Compute focused on more ecosystem focused
- Stay tuned for collaborate site
- HPC Needs Assessment
    - Document sent out to all
    - Assessment of what we need across CBIIT
    - By end of this month as it stands – what do we need to change in terms of making sure we have needs identified that aren't there and taking things off list that aren't needs
    - Feel free to edit accordingly
    - Needs assessment will dovetail into feeding LRP and inform priorities for FY17
    - Needs assessment gives a broader scope
    - Try to collect info about HPC needs across CCR, DCEG, and other places to address (Sean and Dianna)
    - Greg: understand if there are compute or storage demands here and Frederick would help with our planning
    - Sean: helpful to go a step higher and have high level decision making about what Frederick is best used for and CBIIT is best used for and what Biowulf is best used for. If we evaluate each year this will change drastically from year to year

DAT services and storage environment

- Greg: Cleversafe update: Cleversafe storage co-investment

We are going through PM change. Made some good progress , operational, working on setting up for select users to do testing

- Working with Steve Fellini and CBIIT about placing nodes there and Biowulf – first set in May and other set in June (June move Cleversafe to shady grove – maybe even July)
- Looking to set up basic training on how to use the system – its not a standard file system. Those interested can contact Greg and invites will be sent to training
- Action: get some of users engaged to get use defined around capabilities and match what is being delivered to a need that is being fulfilled based on use cases. CCR sequencing is one of these groups. Sean to build on that: higher level detail

Steve and Biowulf group are in position to contribute to these conversations. Original implementation plan for Cleversafe in Bethesda is abeer/aver??– direct object storage APIs. If this is the case there is room for teaming up. Solving same problems so should talk as a group to move forward.

Advantages of having object store system is its dispersed geographically. Don't need to back it up.

Nathan – would like to try it out and thinks this is direction to jump on or archiving

Nathan to drop note to Greg and he will set him up with training.

- Eric: engaging in use cases to focus on what services need to do moving forward
- Balancing needs – pushing limit of what technology can do vs extreme reliability and stability (needs for both in environment being served)
- Match program and pediatric match program discussion – they will have some data needs growing substantially (like clinical applications) – match has similar requirements in terms of providing long term data assurance for infrequently used large files (potential opportunity in use cases)
- This is situation where data would likely be held for 5-7 years, but as things would close data will likely move into GDC. Make sure we don't create data resource that is inaccessible to other resources beyond NCI. Out Design with in mind for cooperation without huge technical barriers.
- Greg: next steering committee meeting? March 31st

Other topics of potential interest

- Brainstorming – envision what the future is (Bob Coyne) – if interested in such a session let us know. think about looking ahead

- HPC support efforts (George doing a lot)

  - Globus connect – default tolerance for large data transfer
  - GIT – different ways where its readily usable and challenges with it
  - George some projects working on: request from Dauood, RNE co sammparison taking 5 days for sample. months to finish one analysis. Was running on Biowulf. Original request for space, requirement is 5 terabytes of space. Was asking to get space from one of DOD labs. Must make sure app running efficiently.
    - Application developed by Harvard and California (python). Some of parameter

S are constant. Fixing and modifying to mak1e it run efficiently on Biowulf…from 5 days to run to 1 day. Scaling would do the same thing. This is just one Biowulf node. Parallelism on application is a lot. One node has 32 threads. Challenge is it would require major refactoring of application (not standard). If we go this path we can't support every instance of this application in terms of updates

Not a salable way of spending time.

Name of application is MATS.

MATS NIH – also checked this version today. Currently testing it to see if it fills Dauoud's requirement to run things faster.

Strategy Eric: if its an inefficient application, can we be make it more efficient?

George: its an app written by a post doc or grad student – not requiring much work to adjust

Way to run – always asks for 16 threads, and it only makes use of 4 or 8. Even if asking for more resources for Biowulf, allocated for no reason.

- Home grosn and other applications are being run (mat lab, etc.)

- Open ACC with PGI compiler (now available on Biowulf). If we have more of home grown application, showay and uld have workshop to learn how to optimize their applications in an easy way and making use of our resources in a good way.

- Hackathon in university of Delaware – George going. Load of GPUs and interests – maybe get GPU hakathon at NIH.

- Eric: other things goinhis g on in HPC – support for education and finding avenues where Can be effective

Helping connect those who might have an application with fact that HPC can make more teir research more effective or rapid

Computing and predictive oncology meeting in July 2016

Initiated out of doe collaboration.

Targeted in downtown DC

Working on logistics to confirm venue so we have meeting and define it

It's a short time frame

One of reasons to have it in this ARA is to maximize number of people from NCI to participate with out travel

Around 100 invitees

Pull together where computation is and where its going and impact on predictive oncology

More details come up we will get this out

- Was hoping warren would join us by now for update on NCI DOE collaboration
    - It's definitely moving ahead. Project teams are working together and working logistically to coordinate things across
    - Three pilot areas as mentioned before
    - Molecular biology focus, clinical, population scales
    - Collaboration is taking there these rep areas of predictive oncology and working with DOE to frame where are these problems to be solved in this space to design computers in future to help.
    - Bilateral mutual benefit in that context
    - Things I'll be framed up to include in upcoming Frederick national lab advisory meeting
    - Details of pilots – article in genome web. Interviewed from Livermore
    - What are logistics for each of pilots and cross pilot effort – framed and formed at this time (not a lot to say right now). Typical things about starting a project (funding, teams, support, sharing data, etc.)
    - One common thing across collaborations – all three are excited to work with each other
    - Eric – did listen to testimony in front of committee overseeing NIH budget. National institute for aging and drug abuse. Various things focusing on (NIH pushing for things thy are investing in). Single cell sequencing one thing pushed for. Dr. Lowey – talk about the NCI working with DOE and benefit evident with large computational capability and expertise that DOE has complimenting knowledge and data that NCI has in cancer work.
    - Questions about moonshot and how this applies
    - One takeaway from testimony – essentially gender bias in a lot of studies. Heavily male oriented studies. Study design needs to take gender into account moving forward

Sean: Data management and converged IT – potential conference /Summit

Email thread going out – idea that we have a lot of data and storage and storage needs. Two pieces to that puzzle making it work storage infrastructure and network infrastructure and connecting to that, and second is the meta data. Think about this and if anyone interested follow along in email conversation. Interest in some kind of conference or summit. Invite a few extramural people to tell us how they would have done things and talk about data management strategy at a higher level.

Reach out to Sean with interest.

Creative ideas for funding? CIT might help, warren's office, NCBI.

Commercial sponsorship? Maybe open to that. It's possible but they are not allowed to ask for it. If commercial sponsor came offering to do something we can do it but we can't ask for it. It needs to come from someplace else. Maybe Intel is interested in something like that?

Funds needed to travel extramural folks in and for physical space. Can use local talent but maybe 4-5 ext. folks or potentially someone from bio team.

Maximum number of people expecting in attendance around could be 50-75 people or different approach and have set of speakers more open on first day then more focused group on second day. Could do it either way. Smaller scale could use shady grove (carl). Down staris conference room in shady grove.

Timing is right for this now.

Good time to get everybody together

Conversations continue on slack

<u>Next Meeting – April 21</u>

February 18th 2016

Agenda

- Data and Data management needs
- Real use cases and policy implications
- Establish policy directions through Warren so technology can align to those
    - We have sandbox of technology flexible in place - need to align and prioritize
    - Brainstorming on Data Services aspect
    - Cleversafe update (if Greg joins) – production roll out supposed to be in August. Not most everyone ready to use it
    - Demand much earlier than that point from HPC perspective
        - Planning for Cleversafe to be standard CCR for data storage – 40 labs approx. ready

Updates

- Sean and Eric with CIT last week – saw info on forward looking plan to roll out Cleversafe
- CSS (data coordination center) – using Cleversafe for storage (third party)
- They are getting ready to architect how they would use it – they have software development plan done
- Talking about presenting production level version before Cleversafe is rolled out (their timing is a bit off)
- If its not ready to go with their timeline (don't think they can use cloud version – depend on how they architect it)
- Working to get familiarity with use cases and needs with CCR – limiting step is time to coordinate with production availability of Cleversafe – if it were sooner we'd use it sooner
- Sean Davis: in short run we should consider using Amazon because API is the same – no reason not to architect against amazon for storage in short term
- Let everyone know it's a possibility – we should be architecting for that anyway because Cleversafe is more expensive than Amazon
- Sean provide write-up on how to do that with amazon in short term? (sign up and provide credit card) – put Potts order in. No issues with data security (need to live up to all security requirements). For storage temporary, not a production thing. For testing purposes its fine
- Eric Stahlberg: what type of data can you put up there in general? (anything, just needs to be secured depending on security requirements for that type of data)
- Government Cloud – separate partition of Amazon.
- ***Action item: get a guide from Sean for purposes of designing and developing an app meant to run the Cloud (Sean can't do it).*** There's a work group here at CBIIT involved with getting Amazon services more formally (CBIIT security and CIT security folks on that team).
- For informal testing and development against S3 NCI it is doable for sure
- Speak with Sean first – he's the only one using Amazon right now
- Longer term planning and opportunities to realize

Forward looking direction for Data Management, compute and Cloud (thought s and perspectives) what people would like to do and anticipate they would need to do:

- Nathan Cole: for Amazon, one of biggest things is large scale amputation in terms of taking all GWOS data that CGR collected over years and try to turn it into one master data set and then imputing things across all that. This is not highly storage intensive but computationally intensive. Been able to do locally but shut down everything else longer than anyone wanted. Adding to this data set moving forward and being able to do different things with it in Amazon computing scenario. Little data going up or back but a lot in middle.
- Amputation for data set (over 500 cores) took between local compute and leveraging Biowulf (using at any given time trying to run a few hundred cores) took 4 months from start to finish (total roughly 800 cores). Constricted locally to half available cores due to memory issues.
- Larger nodes would make recombination easier but would not make it go faster. Amount of memory on a node determining how big of a chunk of data you can do
- Compute clusters utilization right now typically is 60-70%
- 4 months for amazon time will be really expensive compared to what we can do locally
- Might make sense for NCI to put in some money for Biowulf strategically
- Biowulf expanding by 120% in next 6 months
- We should do Amazon but there might be a case for taking that money and amortizing it into some local compute
- Do cost analysis of what it would cost to run it on amazon. (Not 4 months straight through of crunching time – lots of other things going on in there as well)
- Local compute used 30% or higher utilization – it's cheaper to do it local (rough cut off)
- If its one-time thing, it's no big deal. If it happens more per year we can't be shut down for that long operationally. Need to find way to spell it out through amazon or other places. Figure out what works best for that. Do testing on amazon and Biowulf and other to figure it out
- How much data to move? Couple of hundred Gigs.
- Impute 2 is name of application
- Lots of options available to balance cost and turnaround time
- Moving data between us and Biowulf or Helix is not a problem but we are looking for better ways to do it
- Couple of request to get Globus set up on CCADs – is there a reason to try and set us up as an end point? (End point and client are the same thing)
- Globus adds authentication and authorization, security key management and cloud automation
- Installing grid FTP server with Globus stuff on top of it once you install. You are installing server with root privileges which allows it to be multi user but underlying technology is the same
- It's a 5-minute installation and simple to do – not a lot to worry about
- ***Sean Davis to send installation notes from Steve Fellini (contact person for CBIIT)***
- Need some ports open (if you're using a strong firewall you'll need to open some ports)
- Put it on a system that has as much bandwidth as possible and bypass the firewall
- Do fairly early on – talk to science DMZ folks
    - ***Sean Davis will send some information to everyone***
    - Jack Collins: talked with Java folks – they need place to back up their data
    - Eric Stahlberg and Sean Davis spoke to Anjan
    - Will have to continue to push them to use technologies already available. Reticent to use Globus but will have to bite bullet and do it. They will be one of Cleversafe users. Process to onboard them. Change way they work a little bit.
    - Cleversafe not up and running right now – projection is August
    - In meantime – we have storage for them 3-400 terabytes at CBIIT and can get more.

- ***Issue: they have 400 T on new system and they want to back it up. Back-Up system that CBIIT uses not appropriate. Instead mirror data over to Isolon where they could more appropriately manage backups***
- ***Sean Davis will install things for them and get them up and running and make sure storage is available for them***
- Keep everyone in the loop and make sure they are ok
- What timeframe are they anticipating producing data out of Clenomics efforts? Summer 50-100 patients in first year (1 T of data – nothing serious). Issue with them is they tend to spread out. Had 400 available and before project started they had 120 on it. Need to figure out (cant keep making copies of data). Data management plan is Clenomics is currently stored on 4 copies.
- Its not amount of data but Data Governance (discipline)
- They are bringing on 10 new people – hoping we can task one of them with this
- **If groups want to use Cleversafe as a parking place we have to have a business model to do that.** What is business model to sustain Cleversafe as it moves into production. Have good list of use cases. Place where we put non changing data – have copy (safety factor). Not built for transactional data. Clinomics data versus other data? How do we figure this out? Not wise to put data on Isolon – should be able to explode it. Need to make changes on how we take this data (until we have longer term plan) – understand use cases. Band aid solutions are not good for any of us long-term.
- Make sure not rebacking up data in Isolon (static data – check with Jeff and Javed's group)
    - Being able to utilize this data and use it in other studies, retrospective analysis – is to have appropriate meta data to be able to do that
    - Making data valuable for use of researchers just as important as having place to store it
    - Any of these solutions include meta data (Sean Davis)
    - No single meta data solution that will work for all projects
    - Different approaches for different groups
    - Educating groups and provide minimal support services to help them create solutions – do in way fairly local to problem – when we have local problems that have global solutions we can apply those
    - Need to figure out how to incorporate it for each lab or branch. How to globally control NCI data (this thinking won't get us very far)
        - Some use cases may just need place to park data, others for meta data
        - Need to figure out how to triage it
        - Sean Davis: sit down and discuss those things – ask labs what their use cases are (not valuable way to go about things)
        - Architecture level to look at it – here's the way to connect things
        - We have idea of types of data and requirements levels currently
        - Having a service catalog like list of use cases and triage it in – not everything will be perfect fit (some will be shoe horned in)
        - Push technology and context of use cases 80% API covering most use cases.
        - Don't try to define one size fits all (evolution and uniqueness to meta data)

Keep base technology as general as possible so we don't put ourselves in a corner that's heavy lift to transition

- Get people in lab used to following some kind of best practices for discipline
    - Having a steering committee monitoring some of projects going (WebEx and HPC data initiative). Communications between groups not always seen. SC can facilitate communications and align projects together and can address the other issues tactically and operationally
    - Put items on list of collective group to address and focus on
    - ***Sean Davis: another strategic item: some of reason we struggle is that we don't have right bodies to do some of this work. Worth thinking of what kind of bodies we need. Relevant to get somebody in NCI who has some expertise in these areas because none of us do (big data management and cloud infrastructure). We can get it by learning but we are in a place where getting someone in with experience will save us a year's worth of work.***
    - **CDC contracted with SRA**
    - **Bioteam working with USDA and a couple of Universities to do this**
    - We should find someone potentially to help move us along more quickly
    - Educating people- we have a lot to do there (require some investment from CCR and DCEG) – scientists learning some of this stuff – best done within scientific program and not within CBIIT right now
    - Jack Collins: Bringing in good people and setting up test lab of people who know the hardware and people supporting scientific applications from a system administration point of view at this cutting edge places
    - Nice to have some hands on experience with some of these technologies (even if vendor comes in and we have 3-4 people and translating that into our day to day practice)
    - ***Will require some funding for FY17 (Sean Davis working on funding side)***
    - ***Jack Collins to scope size projected FTEs would be or budget number would be around 400-500 direct (ballpark)***
    - Need to figure out as a group how to get it to exist within NCI

    - Sean Davis: my idea is similar, but we need to have these kinds of people directly tied to the science (physically sitting next to scientists) – have seen it a lot. Divides currently exist. Way to move this forward is to bring the IT into the scientific programs rather than something separate that supports the programs.
    - When get people primarily focused on science at CBIIT then transition can begin.
    - Data engineering problems within intramural program – we need something agile right now
    - We need both? Agile and someone to put infrastructure together that meets it in the end. Should keep both programs going and connected at the hip and talking to each other to know what's going on.
    - Group building infrastructure is closely monitoring what's going on out there.
    - ***Caveat: much of knowledge about infrastructure is embedded within scientific program and not in IT – will change over time. Cutting technologies easier to see from IT side rather than scientific side right now.***

Next HPC Thought Leaders Meeting – March 17, 2016

January 21, 2016

Agenda

- General Updates
- Storage and Data Management
- Long Range Planning
- DOE Collaboration
- FY17 Priorities

Quick Updates

- Cleversafe – Greg Warth: Equipment is here and staff spoke with Cleversafe. They are coming on site to discuss installation issues. Expect installation and up and running by middle of February. Hoping we will have it so people can start using it in test mode sometime in March. Writing project plan and putting dates in there right now. Need to know level of protection we are looking for in the system (how many nodes to make on there), and when to expect to do limited and full production. Will be in contact to ask more detailed questions. First part of installation – all 4 nodes in Frederick, then nodes in Shady Grove and one in Biowulf.
- Nathan Cole: No new updates. Want to get in the loop with regards to Cloud Pilot – has not heard anything. Has anything been made available? (Sean Davis will give update on that)
- Xinyu Wen: no updates. Everything is working fine. Negotiating for storage – in progress
- Dianna Kelly: no updates
- Jim Cherry: not on line
- Steve Fellini: no updates
- Sean Davis: quick update on Cloud Pilots:
    - All 3 are officially open for business. Seven bridges, ISP and Google are open (4 and 7 weeks) Brode opened yesterday. 7 bridges Is pretty robust – instantiation of their cloud platform with bells and whistles and access to DCG data
    - ISP google is rough around the edges. One piece working well is sequel access data base (big query) loaded most TCGE data in to that system and meta data (clinical and file level). Workflow management system and data management is still up in the air. Web portal is essentially in alpha testing.
    - Brode- nobody seen anything but screen shots. Open for business but way they are doing things means they are accepting applications but will take couple of weeks for people to get access. No additional detail.
    - Carl McCabe: Intramural Retreat – did you come across new people with ideas you were not aware of?
        - Sean Davis: CCR chunk of Biowulf – many people didn't realize that resource existed or what problems they address. It's a problem we'll have – need to know how to communicate to the right people. Human resource level to devote to these projects. Question of how much more resourcing we should put in to make sure infrastructure element is maximally utilized. We don't do much at this point. Need strategic planning for storage and data management. Needs to include human resources.

- Representation from three different groups said they need to move data and how? Same thing with compute – our scientific computing needs (genomics, etc.) are being met, but we have needs that could be addressed using HPC that have not been traditionally addressed that way.

- Grid FTP server support in Frederick? Storage and data management piece: need resources to back up what we stand up. Sean hesitant to publicize Frederick instance due to not knowing life cycle. Question also of communicating once we set these things up – what are implied or explicit policies in place and what are they meant to support?
- Issue isn't using Grid FTP but making sure we know what happens with the data. Data transfer is convenient mechanism but users need to understand where their data needs to go.
- Moving forward need to think of data management and storage as a suite we offer and need to be careful that Frederick, intramural, etc. are all on the same page. Need to have the communication to support the technology that is already in place.

- Eric Stahlberg
    - familiarity and education challenge to raise awareness of what we have and how it fits their needs. What is there and how far can we extend resources to data management and storage and compute. Not overpromise and under deliver or under-perform.
    - General request mechanism in place. Eric and George are filling up back-end. This is only the triage part – needs to extend out. Bringing forward meta data and upgrading.
    - Have submitted a couple of requests. People contact us by email, etc. We are guiding a submission into that so we can get history and profile of types of request coming in.
    - Differentiating informal requests from formal ones that need action. – raising visibility there to get resources and use in future planning. We have pieces in place. Now we must move forward with the execution. We had talks about different roles and functions. Now we have the pieces in place.
    - Jim Cherry: no updates

Storage and Data Management

Make sure we start talking through the issues.

- Have services for moving data, soon will have for archiving data
- How do we push forward to meet today's needs and needs of future?
- Must have what people need and want in terms of data storage management capabilities
- Globus Grid FTP (Sean Davis) – put on agenda- we have two working systems and Biowulf Helix is third. We have nice data fabric including Frederick , Shady Grove and 12A – should invest small amount of Human resources volunteering our time to talk to people (didactic) – show people what we can do with this
- Request generated at Duke in sequencing – how do we access our data? – Sean showed how to do that with Globus. This kind of thing would be helpful in pulling people out of the woodwork. Asking question how can we help you doesn't elicit answer – showing technology does
- Carl McCabe: put something on NCS website – illustrated walk through.
- Sean Davis: hands on, tutorial approach (not complicated technology). People may shy away if they haven't seen it done and think its' complicated. It's not hard to use once you've used it once. Initially learning curve is daunting even through it shouldn't be. A lot of opaque concepts.
- Sean Davis: how to deploy Cleversafe? Details slightly above Greg's questions (how to carve up storage for data integrity) , but more along lines of usability. Specific ideas on how to use system that may not jive with what people are thinking. Want to think of Object storage as object storage having people access data using as 3 swift or Cleversafe FDI directly, as opposed to having light or heavy duty front ends on that. There are many third party tools, software and documentations built around S3 and Swift. If we don't expose these APIs we will end up having to create infrastructure that we shouldn't have to create. Also, moving NCI forward in how we think about data architecture, engineering and science. Try to adopt technologies that allow us to leverage our data wherever it sits.
- Should be able to access data sitting at S3 on amazon from Biowulf in same way access data sitting in Cleversafe system within NIH from within an instance on amazon. Important because there will come a time not to far out where we need to use resources above and beyond what we supply locally. Should not box our selves in to set up whole new infrastructure for our data management. Nice to have Amazon access S3 from Cleversafe in same way we access S3 data from Biowulf (buzz word touch all for cloud storage).
- Greg, Jeff and Dianna doing long range planning for storage – should meet to go over different things we expect storage to do and interfaces we need to build, APIS to expose and how storage will work across all NCI on it.
- Meeting to go over Cleversafe is a start (need to capture all this). Unified storage model – adapt what you're trying to do here. Should be able to build in.
- Need to involve CBIIT, ISP, Frederick, CIT folks to incorporate Cleversafe
- Sean Davis: meeting in Shady Grove to allow Biowulf to see data center, then meeting at 12A to allow CBIIT and Frederick folks meet and greet and see the CIT data center.
- Mid February before anything is installed and march before anything is operational – want to do it right rather than quickly.
- Sean Davis: API's and user interfaces are same thing, not separate. If you do expose SP or swift API there are clients, you can download from internet that immediately access the S3 or swift based storage and can mount it on local system. Leveraging software that already exists rather than write stuff to deal with our data management needs.
- Eric Stahlberg: There is whole point of accessing data set and computing on that data set. All these pieces come into place for performance. Find ways to make that as transparent as possible so there is persistence, supporting innovation and analysis.

- ***Greg Warth action item: put together a meeting.*** Sean suggests agenda item – give an overview of scientific work flows and data management as its practiced in the field right now so we're starting from same scientific understanding – can get in weeds about how to implement things. We need to be all on one page from understanding – potential set of use cases.

- Eric Stahlberg: Cloud Pilot provides set of good examples to look at (data management, object identification, etc.) concept of universal resource identifier for data sets. Across all instances that is a common need. One of core elements of strategy is how are we going to satisfy that need? Enterprise decision that transcends NCI

- Where do we find long term archive storage? Critical need because amount we have sitting on expensive disk. Sean has to do to get back together with Allison from Microsoft. When he does, he will do it as globally as possible. Need to find place to put this data.

- ***Sean to send communication to everyone who is at this meeting and Greg to organize a meeting. Will try and pull things together about Globus connect and some example illustrate cases of what to do. Get together with Microsoft to move data from spinning disk to more affordable resource. Will discuss meta data and data sharing within context of that meeting***

HPC Long Range Planning

Eric to send document that was pulled together – this will get informed and updated so people can take a look at what was done.

- Document talks to some projects aligning with storage and data management (more data management) to support compute and HPC and avenues to secure resources on and off site (large scale centers and efforts to go into cloud). It all converges and pieces are coming together. Needs are difficult to meet independently.

- **NCI-DOE update:**
  - No follow up meeting happened because there was no context for it yet
  - What's happening: NCI has opportunity and need to pursue innovations in different areas (understanding biology of cancer), developing clinical trials and understanding more completely impact of cancer and treatments on population overall. These are areas of pilot explorations which overlapped well with needs of DOE – advance Exascale computing, in context of National Strategic Computing Initiative (cross government effort to build more computational expertise not in silos, Government and Private in effort involved).
  - NCI and DOE collaborated gave overview of pilots that would be pursued – agreement was that effort should go forward
  - ***Kick off meetings are being targeted for Feb. 1-2 : context is to put in detailed plan for those focused pilot activities. Overall implications of how they extend and where they go is TBD.***
  - Aims and goals have been laid out but detailed plans have yet to be defined (context for kickoff meeting)
  - More communications being developed about this to go out. Contact Eric Stahlberg or Warren Kibbe with any question

HPC FY 17 Priorities

- Not identified yet. Discussions and meetings going forward will help shape what we prioritize for FY17.
- Focusing on DOE right now for communications plan (talking with Shea, but doing long term perspective to pull things together – web pages updated and talking to people and getting word out)
- Formal presentation to Shady Grove, Frederick et. all – must do formal presentations as part of communication campaign (in 37 depending on size). Cannot expect people to understand anything from emails or just collaborate. Need contact in each building to get a room and start talking to people directly.

Next HPC Thought Leaders Meeting - February 18th, 2016

December 17, 2015

Pre-Meeting Discussion

- PM collaboration effort
- Project management support so we have shared area for various documents amongst team members and add team members to projects as we need them
- Service now could meet our needs but needs to be expanded beyond current operational focus

Agenda

- DOE Collaboration
- HPC Long Range Plan Feedback

DOE Collaboration

- Long range plan not completely guided by DOE collaboration but there is some input and feedback
- 15 people in room from NCI and DOE (members and leadership), national labs, Livermore, etc.
- Speak in context of national strategic computing initiative, Precision medicine initiative
- Proposed collaboration intended to support both initiatives
- Three pilots
    - Predictive algorithm for cancer therapy
    - Predictive models for pre clinical screening
    - Multi scale data modeling using RAS initiative
    - RAS proteins in membranes
    - Milestones for 3-year pilots

Questions

- If things move forward how long will this go on with DOE and how will we phase out?
    - This is going to go forward at least in terms of sufficient support for planning to go ahead. Expectation is that planning will meet the bar to proceed
    - In long term what will NCI do? this is a pilot that clarifies role of scientific computing in cancer. Not only thing to consider but helps us clarify what we can expect to deliver
    - What is NCI thought on integrating people already here on some of these other efforts?
        - Start broadening teams and broadening involvement in various pilots
        - Learn more about pilots and be more involved and transition expertise that leads away from being fully dependent on DOE for these pilots but being able to carry it forward
        - These people don't have expertise in biology or oncology – mistakes will be made in Exoscale if they are made
            - Get an idea of NCI level support and what it will be and support those integrations
            - Co-location – scientists from DOE collocate with Fredrick scientists to collaborate and share ideas. Specifics not yet in play but needs to in order to go forward
            - This group needs to get engaged to go forward
            - How about Frederick scientists engaging with Frederick scientists? Breaking down silos within NCI not just between DOE and NCI
            - Lots of interactions between various groups in piecemeal
            - RAS part and EM images – a little interaction, and some touch points, we have something to bring to the table if the table were offered.
            - Need to have workshops to get people to come together that have shared interests to talk and communicate
            - Need to have logistical support with which to follow through
            - NCI needs to be the one driving the science – has a lot to say as to what is the most effective way to move the science forward
            - Where does location of physical system need to be? Argon?
            - What is it that would need to be carried forward from the pilots back to NCI and how? Need to have Exoscale capacity and how to deliver that? Need to have access to NCI remotely or here?
            - We should focus on the science – DOE and Exoscale should be irrelevant. If coral is only place we can do it then so be it.
            - Tempting to think because its an HPC pilot that we need to use newest technology – it ultimately needs to be sustainable and portable for us to use
            - Have to be careful what the priority is – ***priority to NCI has nothing to do with establishing Exoscale in biology – NCI cares that we get science out of this.***
            - ***Our visions need to be aligned with DOE***
            - Needs to not only be focused on science but also scientists as well – whatever is developed needs to be able to come back and help the people in the labs and clinics at NCI do their jobs better
            - Need to flesh out NCI goals when we plan this (write it out as a grant). Technology piece is one aim but not a primary aim for the pilot
            - Whether this impacts cancer or not we don't know. But we are posing our challenges to the DOE and what we get out of it / what NCI is putting into it is sharing the information in the domain
            - Guidance of what works and what doesn't is important for them to not spend their time on dead ends
            - How many FTEs? Discussion between warren and Doug
            - What do warren and Dr. Ishkal want from us the HPC group?

- Eric has been driving this – without him it won't go forward
- They want to get the science done
- We help them get science done in near time and sustain ability to do the science in the long term (this group). High level context to advance the science and precision oncology
- Structure around data and interfaces? Expecting anything form CBIIT, IFOG to set this up?
    - Should be smart observers of what is going on so as things develop and are promising we use that to go forward. Be in a position to facilitate the involvement, not just getting data into DOE but getting our scientists plugged into the projects
    - Set up procedures that will work over broad range of things
    - We need to be able to set up and a group of people driving it not just one (be broader)
    - As we move into detailed planning phase
    - Machine learning tools, machine learning at large scale, … learning
    - This can be one giant learning of what doesn't work in all those three areas

- How do we support Eric?
    - Having discussion now is informative
    - Pull together group to say what do we need to do to support this near and long term?
    - Skeptical optimist…
    - We need to have extramural people involved (computational biologists, machine learning, natural language processing people) – pull them from academic community to be our go to people, not necessarily doers
    - Is there any new money to bring on people?
        - If we don't develop some expertise internally (building staff, etc.) we will be in the same position in three years
        - To build team wherever we do, we need leadership who has done this before so we know what kind of people to hire (give us names, approaches that can work), there will be knowledge transfer
        - We need to have people to council us and we need to be dedicated so we have control over it and not lose it once projects are done
        - (all agree)
        - Resources that are dedicated to this and interact with other groups at NCI
        - We shouldn't wait until they are over to implement transfer of successful things when we identify them
        - Workshop in super computing is first step to build this effort to augment that and begin with it
        - ***Need dedicated resources (not IPAS or detailing from other parts of Government) potentially contracting from data science companies, need money set aside for this***
        - We have to scale now from our resources (lets' star putting people in place to peer with DOE and move forward) so we can learn
        - ***<u>Identifying who we think thought leaders are and start pulling them in</u>***
        - Identify specific timeframe milestones
        - Intention of meeting was to present to secretary of energy to see if this was worthwhile to move ahead with – agreement was yes let's move ahead. Scope and budget are being discussed right now
        - Determine whether this will have impact on cancer research that is done computationally
        - Feasibility is also discussed in this
            - NCI – have you advanced understanding of science
            - Need to treat this like a grant, not a pilot. 75% should be guaranteed to be done. 25-30% can be high risk
            - We need to define the 75%
            - What can we do near term, what is concrete? Then what is the skeptical stuff that we can do afterwards? Some of it might fail
            - One role of this group is that's these ideas role forward they get sanity checked. Right eyes from NCI. Not just happening in a silo or a back room
            - We need to do something other than an hour long meeting each month.
            - Need a different tact and approach
            - We need to be more concrete. Oversight or expertise that doesn't yet exist we need to make sure gets put in place early. Identify high and low risk things, have right mix and resource them appropriately and according

<u>Next Steps and Action Items:</u>

- Getting group together
- Have another focused meeting to discuss in greater detail what we need to do
- Distributing what information is

- Getting involved in the details
- From NCI side of things, people involved at the table aside from the pilots
- Grants management and contracting, admin and infrastructure and computational – who are the leaders?
    - Warren, Dave Heinberg, Dwight, etc.
    - Supporting warren and working with Doug
    - Julie Klemm
    - What is extramural aspect of it?
    - Group is very high level – not a lot of science "doers" or reality checkers

- Must have meeting to decide how this project is going and what needs to be done
- Need steering committee on the NCI side
- Dangerous to say DOE will take care of the IT. They need to tell us exactly what they are doing so we can fill gaps and provide resources
- <u>Steering committee with sub working groups – IT, NCI admin, computational, computational biology – don't worry about scientific side of things</u>
- Need to find these experts somewhere
- Do it before we get much further along
    - Use org chart to try to figure out what fits

Next HPC Thought Leaders Meeting – January 21st, 2016

October 15, 2015

- Attendees: Jim Cherry, Greg Warth, Sean Davis, Steve Fellini, Diana Kelly, Omar Rawi, Rahcyne Omatete, Eric Stahlberg, Kelly Lawhead, Carl McCabe

- **Announcements**
    - Third Thursday next month is week before thanksgiving (super computing conference– next HPC thought leaders meeting will be held in December as opposed to November)
    - We are on track with what we have for Long Range Planning – putting content into template and sharing that around to get input

- **HPC Resource Updates**
    - There have been some significant investments and commitments made since these meetings started. Anchor points in place to form longer term strategy and how best to use those
    - CBIIT Globus connect server update – purpose has evolved to being pre archive limited term data storage needs (people don't know where to put data), now they do have a place for a limited term while long term destination is being identified (requesting storage etc.) Serves as a storage shock absorber for those with large data needs. Operational status is: it is in operation, backups being made, limited retention while duration not specific necessarily, people understand it is for a limited time until move to permanent location. Global share has been enabled, and server functionality has been tested.
    - Started to get open requests on what this server can do for us.
    - Comment: Useful to go through storage inventory that CBIIT has and identify large scale storage users and come back around to them to suggest a potential option to change that network mounted drive concept to more of a storage end point (CBIIT and Fredrick) – critical strategic piece of removing desktops from large scale data movement (making sure that does happen)

- **Cleversafe Update (Greg Warth)**
    - Decided to change backend storage and there is a delay. Expect to have it in mid November and will install in Fredrick (around 1.2 Tera bytes of usable storage). Plan on putting one of nodes at Shady Grove, and talked about possibility of getting one of them at Biowulf (still interested in doing that)
    - Need to talk to someone to get permission for data center (talked with CIT and network group to get network set up – then we will talk with data center folks to get it done). Identify contact person for data center (Greg will identify someone)
    - Eric: who is sharing in that capability from Cleversafe? Sean is majority of funding provided but also Eric and Andrew quam with data storage repository funded the project
    - We will look to see (set money aside) if need to put money in front of it (CCR and CBIIT put money into it to get it up and running and Andrew qualm's group)
    - *CBIIT has long term strategic planning process (long range planning) but storage will go online long before that is concluded – need to understand what our business model for storage is going to be. Discuss data policies and include CBIIT storage group, Biowulf, DCEG potentially and ISC to understand where we are headed and what the*

guiding principles are. Action Item: Get a meeting with this group to start talking about those things

- **Cloud Resources Update – Sean Davis and Azure specifically:**
  - Azure: CIT signed extension to enterprise agreement with Microsoft to give access to services (O365 and MS 1 drive – drop box like solution per user) for storage and other cloud services Microsoft has set up an accounting system through CIT and CIT is figuring out what that is going to look like and how to charge institutes and take care of billing for cloud services happening outside of Microsoft 365 and one drive
  - Will be an annual basis with a start date and contract will continue through calendar year – during which can park money in a CIT account – money can be used to charge for Microsoft services. End of year, account needs to be at or near zero because money doesn't roll over into new year
  - How this will be done is totally up in the air right now but is being discussed – need to have meeting with CIT when farther along so at NCI level we know what central management needs and groups to access services will look like.
  - Trying to find archive storage service so need to see what pricing model is and how we can charge so it will be advantageous to have that discussion
  - Storage model similar to amazon storage model – they don't have anything equivalent to glacier or near line
  - They offer piece on store simple – similar in concept to azure gateway (Good way to cross migrate some of data users)
  - Unlike amazon cloud where you need to migrate things in HGSS based system – Microsoft direct storage in cloud (opens up potential uses)
  - Need to wait for CIT to sort out details of billing – when they are ready we will meet with them – we do have a Microsoft representative on campus who is technically savvy ***(Talk about this at some point later)***
  - AWS (Sean) what were some of your experiences: had access to amazon for 6-8 months – using it relatively regularly to try out. Set up servers to try out databases, analytical approach implementations. Tracking project management system.
  - Did not have to download large data – have identical environment for 20 students, worked really well in academic setting
  - Discussion – NIEAD – Landa services – automated run on certain triggers
  - Azure and AWs are two cloud resources we can get to: there is some precedent now we can turn to in terms of what may be possible approaches when we need something. We have something to build from
  - Focus on azure for the short term but AWS is so much more ingrained in research computing – we need to target that. Talked with CIT.

- **Biowulf 2 update - Steve Fellini**
  - phase 2 contract a week ago comprised of 30000 cores, moving to a EDR backbone, getting couple of additional DDN storage systems, 4 petabytes, requirements of 4 Gb per second bandwidth.
  - Electrical work in data center will delay installation (hoping for march but its up in the air right now)
  - Capacity increase from 28000 cores (good number of those are older slightly technology – I gigabit per second rather than 10). Processing power still decent
  - Will be retiring several thousand of the cores. Not quite doubling capacity
  - **Data Management Service Update – Eric Stahlberg**
    - Phase 1 earlier this year – aggregating data and meta data and figure out ways to deliver data services to users
    - Project extended couple of months because of developments with Cleversafe – effort established operational capability to take large file with meta data using Globus server. Defined interfaces most useful for users
    - Decision to move ahead with Cleversafe was taken – extended by couple of months so team can look at Cleversafe and experience instances mapped as we've defined them in the cloud
    - Resource map to use cases as we've defined them is being done
    - Looking at product like Irods as infrastructure to manage across data sources – team look at what it would take to use Irods as bulk environment. And mapping user needs to that
    - Instrumentation introduced between resource and user to understand needs, bottlenecks etc.
    - Understand what phase 2 would be – take different pieces over last year and know interfaces and use case and meta data interest level and pulling that into phase 2 to stich together these various resources

HPC Needs Review and Refresh

- Support for long term planning – focus for coming year
- People, storage, data management support, cloud access, more cycles (needs for coming year)

Comments from everyone on priorities and order

- Plan to manage data, not just storage but right storage (we have expensive storage that should be archived – but having idea that we need life cycle management of storage is very important)

- Kelly – outreach and customer engagement so people know about this and what we are doing
- Potentially putting something on service now (Nikola) to begin to support HPC needs through that interface
- Requests for support are pretty diverse now there is no pattern

FY16 HPC priorities

- Education, outreach and training
- NCI –DOE collaboration
- Data management support – not just storage but making sure we manage the data
- Cloud explorations – how best to use the cloud for what we need to do. What is it that we should do there and how? Azure? What's the support model and business model?
- Leverage new HPC related investments – effectively streamlining things that CIT has put in place – making sure infrastructure begins to emerge
- HPC support processes – establishing this and making sure as services and capabilities are there we have a good way to handle enquiries and requests and deliver
- HPC analyst starting soon – background check not completed so will be on day by day basis as to when start date will be
- Make sure there are no requests that get dropped through the cracks

-

Other items

- November 10th invite - Carry over items
- Spring summit in DC area
- Website for HPC efforts
- Call out IT strategic plan since priorities came out of that (to maintain alignment)
- What are mutual organizational benefits for NCI and DOE to working together? Set up special time to run through view graphs that have been developed.
- Three pilot areas identified as NCI priorities
- Supporting the RAS mission – developing higher end larger scale computational models with higher fidelity to explain interactions of RAS interacting with membrane of cell
- Pre clinical models
- Integrating broader clinical data

Eric action item – get input from NCI: Get transcript from last presentation that Warren did (transcript of questions to see if it can be better developed)