MSRI data processing outlook
L. Blackburn, 2019.02.01

The MSRI includes expansions to the EHT

- Additional stations, doubling of N
- Wide bandwidth, e.g. up to 128 Gbps, dual frequency
- Wider bands from faster samplers, from 2 GHz to e.g. 8 GHz

Data processing challenges

- We have reached the **limits of HOPS** calibration for EHT,
    - **No amplitude calibration**, existing band-aid solutions to amplitude calibrate with custom ehtim-derived post-processing, significant source of unreliability in our data products largely missed by data reviews and validation.
    - No **unified handling of gain calibration**, ran out of control codes to tune phases, inconsistent gain parametrization (syntax) for different systematics makes it very difficult to properly calibrate in a global sense, every calibration parameter is a “one-off” effect with a special syntax.
    - Limited ability to phase calibrate on **subchannel resolution**
    - Adhoc phases are not **polarization dependent**
    - Limited **documentation** for data formats and procedures, significant amount of uncertainty about **what is being calculated and recorded**
    - **Outdated data formats** not appropriate for global calibration and not easily integrated with modern programming environments
    - Cannot correct for **polarization leakage**
    - Cannot continue investing into classic HOPS: little buy-in from rest of collaboration, very few experts, and no efforts to unify framework with other parts of the analysis chain. **Work on classic HOPS necessarily distances calibration from imaging and modeling**, when they should be coming closer together.
- Wider bandwidth and larger array for **increased data volume and computational cost**
    - Need to move to **distributed computing** for processing
    - Modern data format and environment must be parallelizable (CASA has significant problems here)
    - As we move to high-SNR regime, ability to **model and calibrate systematics** becomes the limiting factor for image fidelity. The **complexity of systematics** (model of the instrument and environment) dwarfs the complexity of the image (model of the source) itself.
- **Synergy with efforts in Geodesy**
    - Already demonstrated by the **success of adopting HOPS for EHT** work. The **precision** obtained by HOPS in measuring delays, correlation coefficients, and their associated **statistical errors** far surpasses that of AIPS and CASA.
    - The correlation centers (MIT Haystack, MPIfR) are also **major centers for geodetic work**, with associated resources and expertise.

    - Anticipate that continued co-development of VLBI data processing infrastructure will be mutually beneficial, specifically:
        - **Data formats:** Design modern data format for storing visibility data and associated meta-data, appropriate for tools used by geodesy and EHT ("HOPS2"). Leverage the experience from design of the Measurement Set, but make sure format is easily used outside of CASA, and appropriate for parallel-computing environment. Currently the EHT suffers from different data formats at every stage of the processing pipeline. For example there are four conceptually identical data products with different formats (Mark4 corel, HOPS fringe files, UVFITS product, alist summary)
        - **Atmospheric modeling:** Atmopsheric phase is the dominant systematic with respect to fringe detection and the reduction of statisitical error. The simple algorithm used for current EHT works well in the case of phased ALMA, but does not extend well to a larger global array without a single key anchor station.
        - **Polarization calibration:** A unified framework for calibration based on the full-Stokes Measurement Equation (amplitude, phase, and polarization) is necessary for proper calibration of the EHT. Many corrections are ideally made upstream in the data, and due to the limitations of HOPS are left to ad-hoc downstream scripts.
        - **Visualization and Validation:** Beyond the fourfit plot. Joint development of data inspection tools (supported by common data formats).
- **Synergy with Imaging and Modeling**
    - **Calibration (HOPS) has been data-driven**, with a calibration framework tuned to solving specific data systematics, e.g. rates, delays, pc_phases, ad-hoc phases are systematic model parameters that are tuned to fit actual visibility data across the array.
    - **Imaging (eht-imaging) has been simulation-driven**, with a source modeling framework that was originally designed to fit into highly idealized simulated visibilities.
    - Both have moved closer to the middle, with calibration adopting a global framework and modest source assumptions (compactness, variability, circular polarization, etc), and imaging adoping limited gain parametrization for self-calibration. But the **difference in provenance** shows.
    - We should unify the **data format/description**, **gain model**, and **forward simulation engine** across calibration, imaging, and model fitting.