Distributed Correlation for VGOS observations

Simone Bernhart, Reichert GmbH/BKG
Laura La Porta, Reichert GmbH/BKG
Ruediger Haas, Onsala
Stuart Weston, Tim Natusch, Sergey Gulyaev, Warkworth
Jamie McCallum, Hobart
Fengchun Shu, Shanghai
Jakob Franz Gruber, Johannes Boehm, TU Vienna
Gino Tuccari, MPIfR/INAF

7th International VLBI Technology Workshop, November 2018, Krabi, Thailand
Outline

- VGOS - VLBI Global Observing System
- Distributed Correlation (DC): R1840 - a Pilot Study
- VGOS today - state of the art
- Pros and Cons of DC
- Conclusions and Prospects
VGOS

- VGOS - successor of VLBI2010 (IVS WG3) - renamed in 2012:
  - Goals:
    - 1-mm position accuracy on global baselines
    - continuous measurements for time series of station positions and Earth orientation parameters
    - turnaround time to initial geodetic results of less than 24 hrs
  - Requirements:
    - New observing system based on small antennas (12 - 13m diameter), fast-moving
    - Broad-band receivers (2 - 14 GHz, four bands)
    - Recording rates of 8, 16, 32 Gbps
  - Observing strategy:
    - Constant observation with 16 to 32 station network
    - One observation every 30 s

Projected IVS Network in 2020 with VGOS stations
Distributed Correlation

• Aim at testing distributed correlation for future VGOS sessions

• Each correlator only receives the raw data for part of the session (provided that one day corresponds to one session).

• Possible scenarios:
  
  • Main correlator + "branch" correlators, where main correlator does fringe search, preparation of vex and v2d files for correlation, post-processing and database creation

  • Branch correlators only: correlate and post-process the data, upload databases for analysts

• Prerequisite: all correlators use the same DiFX and HOPS (Haystack Observatory Postprocessing System) version for correlation and post-processing
The Bonn HPC Cluster

DiFX software correlator (Deller et al. 2011)

- 68 nodes x 20 compute cores = 1360 cores => 10 x higher computing power w.r.t. old cluster
- 3 head nodes => possible to run more correlations in parallel
- 56 Gbps Infiniband interconnect between nodes
- Storage space > 1 PB, organized in BeeGFS file system
- 2 x 1 Gpbs Internet connection

- 14 Mark-5 playback units
- 8 Mark-6 units with 4 bays
Distributed Correlation

- Geodetic experiments currently processed with DiFX-2.5.2 (difx output - Swinburne files) and HOPS v3.18 (Mark4 format), conversion between the two data formats done by difx2mark4

- Distributed test correlations of R1 (rapid) experiments, bi-weekly sessions (R1 + R4), EOP results on a timely basis, S/X, data format 512-16-2 (256-16-1)

- First attempt of distributed correlation performed in 2016 for R1785 (A. Bertarini) - inconclusive

- **R1840**: 2 May 2018, 122-1700 to 123-1700 (doy + UT),
  - Participating stations: Ht, Is, Ke, Kk, Kv, Ma, Ny, On, Ww, Wz, Yg
  - Setup:
    - Main correlator: Bonn -> vex, v2d file, HOPS station codes and control file for fringe fitting
    - Five 'branch' correlators working on assigned 1-hour time slots
    - Analysis of resulting VGOS database by R. Haas, Onsala
Distributed Correlation

- Data distribution:
  - **Branch Correlator**
  - **Time Slot**
  - Warkworth (Ww) 122-1800 to 122-1900
  - Onsala (On) 122-1900 to 122-2000
  - Hobart (Hb) 122-2000 to 122-2100
  - Shanghai (Sh) 122-2100 to 122-2200
  - Vienna (Vien) 122-2200 to 122-2300

- Data e-transferred to branch correlators (Ma, Kk on module, copied onto raid first)
- After finishing correlation and post-processing, branch correlators uploaded difx output and Mark4 data to main correlator for further processing
Distributed Correlation

- **Comparison of Mark4 output main vs. branch correlators**
- **fourfit statistics (Quality Codes - QC) from aedit (sum 2):**

**Ideal case:**

| Quality code summary for main corr.: | A | B | C | D | E | F | G | H | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ? |
|                                    | 0 | 0 | 0 | 0 | 0 | 19 | 0 | 17 | 0 | 0 | 0 | 0 | 2 | 15 | 27 | 48 | 344 | 0 |
| Earliest scan:                     | 118-122-210107 |
| Latest scan:                       | 118-122-215802 |
| vs.                                 |                            |
| Quality code summary for branch corr.: | A | B | C | D | E | F | G | H | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ? |
|                                    | 0 | 0 | 0 | 0 | 0 | 19 | 0 | 17 | 0 | 0 | 0 | 0 | 2 | 15 | 27 | 48 | 344 | 0 |
| Earliest scan:                     | 118-122-210107 |
| Latest scan:                       | 118-122-215802 |

**QC Code Interpretation:**
- QC = 0: Fringes not detected.
- QC = B: Interpolation error in fourfit.
- QC = D: No data in one or more frequency channels.
- QC = E: Maximum fringe amplitude at edge of SBD, MBD, or rate window.
- QC = F: Fork problem in processing.
- QC = G: Fringe amp in a channel is <.5 times mean amp (only if SNR>20).
- QC = H: Low Phase-cal amplitude in one or more channels.
Distributed Correlation

- Comparison of Mark4 output main vs. branch correlators
- fourfit statistics (Quality Codes - QC) from aedit (sum 2):

**Ideal case:**

<table>
<thead>
<tr>
<th>Quality code summary for main corr.:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D E F G H 0 1 2 3 4 5 6 7 8 9</td>
</tr>
<tr>
<td>0 0 0 0 0 0 19 0 17 0 0 0 0 2 15 27 48 344 0</td>
</tr>
</tbody>
</table>

Earliest scan: 118-122-210107
Latest scan: 118-122-215802

**Worst case:**

<table>
<thead>
<tr>
<th>Quality code summary for main corr.:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D E F G H 0 1 2 3 4 5 6 7 8 9</td>
</tr>
<tr>
<td>0 0 0 0 0 0 11 0 62 0 0 0 0 0 11 23 239 0</td>
</tr>
</tbody>
</table>

Earliest scan: 118-122-190010
Latest scan: 118-122-195846

vs.

<table>
<thead>
<tr>
<th>Quality code summary for branch corr.:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D E F G H 0 1 2 3 4 5 6 7 8 9</td>
</tr>
<tr>
<td>0 0 0 0 0 0 19 0 17 0 0 0 0 2 15 27 48 344 0</td>
</tr>
</tbody>
</table>

Earliest scan: 118-122-210107
Latest scan: 118-122-215802

Ideal case:

- QC = 0: Fringes not detected.
- QC = B: Interpolation error in fourfit.
- QC = D: No data in one or more frequency channels.
- QC = E: Maximum fringe amplitude at edge of SBD, MBD, or rate window.
- QC = F: Fork problem in processing.
- QC = G: Fringe amp in a channel is <.5 times mean amp (only if SNR>20).
- QC = H: Low Phase-cal amplitude in one or more channels.
Distributed Correlation

- Comparison of Mark4 output main vs. branch correlators
- fourfit statistics (Quality Codes - QC) from aedit (sum 2):

**Ideal case:**

<table>
<thead>
<tr>
<th>Quality code summary for main corr.:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D E F G H 0 1 2 3 4 5 6 7 8 9 ?</td>
</tr>
<tr>
<td>0 0 0 0 0 0 19 0 17 0 0 0 0 2 15 27 48 344 0</td>
</tr>
<tr>
<td>Earliest scan: 118-122-210107</td>
</tr>
<tr>
<td>Latest scan: 118-122-215802</td>
</tr>
</tbody>
</table>

**Worst case:**

<table>
<thead>
<tr>
<th>Quality code summary for main corr.:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D E F G H 0 1 2 3 4 5 6 7 8 9 ?</td>
</tr>
<tr>
<td>0 0 0 0 0 0 11 0 62 0 0 0 0 0 11 23 239 0</td>
</tr>
<tr>
<td>Earliest scan: 118-122-190010</td>
</tr>
<tr>
<td>Latest scan: 118-122-195846</td>
</tr>
</tbody>
</table>

**No control file applied, wrong station codes, one filelist contained two stations**
Distributed Correlation

- Comparison of Mark4 output main vs. branch correlators
- fourfit statistics (Quality Codes - QC) from aedit (sum 2):

Ideal case:

<table>
<thead>
<tr>
<th>Quality code summary for main corr.:</th>
<th>QC</th>
<th>Earliest scan:</th>
<th>Latest scan:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D E F G H 0 1 2 3 4 5 6 7 8 9</td>
<td>0 0 0 0 0 0 19 0 17 0 0 0 0 2 15 27 48 344 0</td>
<td>118-122-210107</td>
<td>118-122-215802</td>
</tr>
</tbody>
</table>

vs.

<table>
<thead>
<tr>
<th>Quality code summary for branch corr.:</th>
<th>QC</th>
<th>Earliest scan:</th>
<th>Latest scan:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D E F G H 0 1 2 3 4 5 6 7 8 9</td>
<td>0 0 0 0 0 0 19 0 17 0 0 0 0 2 15 27 48 344 0</td>
<td>118-122-210107</td>
<td>118-122-215802</td>
</tr>
</tbody>
</table>

Afterwards:

<table>
<thead>
<tr>
<th>Quality code summary for main corr.:</th>
<th>QC</th>
<th>Earliest scan:</th>
<th>Latest scan:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D E F G H 0 1 2 3 4 5 6 7 8 9</td>
<td>0 0 0 0 0 0 11 0 62 0 0 0 0 0 11 23 239 0</td>
<td>118-122-190010</td>
<td>118-122-195846</td>
</tr>
</tbody>
</table>

vs.

<table>
<thead>
<tr>
<th>Quality code summary for branch corr.:</th>
<th>QC</th>
<th>Earliest scan:</th>
<th>Latest scan:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D E F G H 0 1 2 3 4 5 6 7 8 9</td>
<td>0 0 0 0 0 0 10 0 62 0 0 0 0 0 11 23 240 0</td>
<td>118-122-190010</td>
<td>118-122-195846</td>
</tr>
</tbody>
</table>
## Distributed Correlation

- Comparison of Mark4 output main vs. branch correlators
- fourfit statistics (Quality Codes – QC) from aedit (sum 2):

### Ideal case:

| Quality code summary for main corr.: | A | B | C | D | E | F | G | H | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ? |
| 0 0 0 0 0 0 19 0 17 0 0 0 0 2 15 27 48 344 0 | | | | | | | | | | | | | | | | | | | | |
| Earliest scan: 118-122-210107  | | | | | | | | | | | | | | | | | | | | |
| Latest scan: 118-122-215802  | | | | | | | | | | | | | | | | | | | | |

vs.

| Quality code summary for branch corr.: | A | B | C | D | E | F | G | H | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ? |
| 0 0 0 0 0 0 19 0 17 0 0 0 0 2 15 27 48 344 0 | | | | | | | | | | | | | | | | | | | | |
| Earliest scan: 118-122-210107  | | | | | | | | | | | | | | | | | | | | |
| Latest scan: 118-122-215802  | | | | | | | | | | | | | | | | | | | | |

### Afterwards:

| Quality code summary for main corr.: | A | B | C | D | E | F | G | H | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ? |
| 0 0 0 0 0 0 11 0 62 0 0 0 0 0 11 23 239 0 | | | | | | | | | | | | | | | | | | | | |
| Earliest scan: 118-122-190010  | | | | | | | | | | | | | | | | | | | | |
| Latest scan: 118-122-195846  | | | | | | | | | | | | | | | | | | | | |

vs.

| Quality code summary for branch corr.: | A | B | C | D | E | F | G | H | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ? |
| 0 0 0 0 0 0 11 0 62 0 0 0 0 0 11 23 240 0 | | | | | | | | | | | | | | | | | | | | |
| Earliest scan: 118-122-190010  | | | | | | | | | | | | | | | | | | | | |
| Latest scan: 118-122-195846  | | | | | | | | | | | | | | | | | | | | |

---

**QC**:
- 0 Fringes not detected.
- 1-9 Fringes detected, no error condition. Higher #, better quality.
- B Interpolation error in fourfit.
- D No data in one or more frequency channels.
- E Maximum fringe amplitude at edge of SBD, MBD, or rate window.
- F Fork problem in processing.
- G Fringe amp in a channel is <.5 times mean amp (only if SNR>20).
- H Low Phase-cal amplitude in one or more channels.
Distributed Correlation

- Comparison of Mark4 output main vs. branch correlators
- fourfit statistics (Quality Codes - QC) from aedit (sum 2):

**Ideal case:**

| A | B | C | D | E | F | G | H | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 9 |
| 0 | 0 | 0 | 0 | 0 | 19 | 0 | 17 | 0 | 0 | 0 | 0 | 2 | 15 | 27 | 48 | 344 | 0 |

Earliest scan: 118-122-210107
Latest scan: 118-122-215802

vs.

Quality code summary for branch corr.:

| A | B | C | D | E | F | G | H | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 9 |
| 0 | 0 | 0 | 0 | 0 | 19 | 0 | 17 | 0 | 0 | 0 | 0 | 2 | 15 | 27 | 48 | 344 | 0 |

Earliest scan: 118-122-210107
Latest scan: 118-122-215802

**Afterwards:**

| A | B | C | D | E | F | G | H | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 9 |
| 0 | 0 | 0 | 0 | 0 | 11 | 0 | 62 | 0 | 0 | 0 | 0 | 0 | 11 | 23 | 239 | 0 |

Earliest scan: 118-122-190010
Latest scan: 118-122-195846

vs.

Quality code summary for branch corr.:

| A | B | C | D | E | F | G | H | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 9 |
| 0 | 0 | 0 | 0 | 0 | 10 | 0 | 62 | 0 | 0 | 0 | 0 | 0 | 11 | 23 | 240 | 0 |

Earliest scan: 118-122-190010
Latest scan: 118-122-195846

QC = 0  Fringes not detected.
= 1-9 Fringes detected, no error condition. Higher #, better quality.
= B Interpolation error in fourfit.
= D No data in one or more frequency channels.
= E Maximum fringe amplitude at edge of SBD, MBD, or rate window.
= F Fork problem in processing.
= G Fringe amp in a channel is <.5 times mean amp (only if SNR>20).
= H Low Phase-cal amplitude in one or more channels.
## Distributed Correlation

- **Comparison of Mark4 output main vs. branch correlators**
- **fourfit statistics (Quality Codes - QC) from aedit (sum 2):**

<table>
<thead>
<tr>
<th>Ideal case:</th>
<th>Afterwards:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality code summary for main corr.:</td>
<td>Quality code summary for main corr.:</td>
</tr>
<tr>
<td>A B C D E F G H 0 1 2 3 4 5 6 7 8 9 ?</td>
<td>A B C D E F G H 0 1 2 3 4 5 6 7 8 9 ?</td>
</tr>
<tr>
<td>0 0 0 0 0 0 19 0 17 0 0 0 0 2 15 27 48 344 0</td>
<td>0 0 0 0 11 0 11 0 62 0 0 0 0 0 11 23 239 0</td>
</tr>
<tr>
<td>Earliest scan: 118-122-210107</td>
<td>Earliest scan: 118-122-190010</td>
</tr>
<tr>
<td>Latest scan: 118-122-215802</td>
<td>Latest scan: 118-122-195846</td>
</tr>
<tr>
<td>vs.</td>
<td>vs.</td>
</tr>
<tr>
<td>Quality code summary for branch corr.:</td>
<td>Quality code summary for branch corr.:</td>
</tr>
<tr>
<td>A B C D E F G H 0 1 2 3 4 5 6 7 8 9 ?</td>
<td>A B C D E F G H 0 1 2 3 4 5 6 7 8 9 ?</td>
</tr>
<tr>
<td>0 0 0 0 0 0 19 0 17 0 0 0 0 2 15 27 48 344 0</td>
<td>0 0 0 0 0 10 0 62 0 0 0 0 0 11 23 240 0</td>
</tr>
<tr>
<td>Earliest scan: 118-122-210107</td>
<td>Earliest scan: 118-122-190010</td>
</tr>
<tr>
<td>Latest scan: 118-122-215802</td>
<td>Latest scan: 118-122-195846</td>
</tr>
</tbody>
</table>

→ data missing in the "good" scan
Distributed Correlation

- X- and S-band observables: total multiband delay (MBD), correlation amplitude/phase, mean visibility amp./phase, residual single band delay (SBD)/MBD, SNR
Distributed Correlation

- X- and S-band observables: total multiband delay (MBD), correlation amplitude/phase, mean visibility amp./phase, residual single band delay (SBD)/MBD, SNR
- Some random plots:
Distributed Correlation

- X- and S-band observables: total multiband delay (MBD), correlation amplitude/phase, mean visibility amp./phase, residual single band delay (SBD)/MBD, SNR
- Some random plots:
  - Bonn-Ww total MBD in X
  - Bonn-On correlation amp. chan. 1 in S
  - Bonn-On correlation phase chan. 8 in X
Distributed Correlation

- X- and S-band observables: total multiband delay (MBD), correlation amplitude/phase, mean visibility amp./phase, residual single band delay (SBD)/MBD, SNR
- Some random plots:
Distributed Correlation

- X- and S-band observables: total multiband delay (MBD), correlation amplitude/phase, mean visibility amp./phase, residual single band delay (SBD)/MBD, SNR
- Some random plots:
  - Bonn-Ww total MBD in X
  - Bonn-On correlation amp. chan. 1 in S
  - Bonn-On correlation phase chan. 8 in X
  - Bonn-Hb total MBD in S
  - Bonn-Sh SNR in X
  - Bonn-Vien mean visibility phase in X
Distributed Correlation

• Comparison of DiFX output files main vs. branch correlator using diffDiFX.py

- Difference on average ≤0.05 %
- Some outliers due to missing data
Distributed Correlation

- Comparison of DiFX output files main vs. branch correlator using diffDiFX.py

- Difference on average ≤0.05 %
- Some outliers due to missing data
Distributed Correlation

- Comparison of DiFX output files main

  Bonn - Ww

- Difference on average ≤0.05%

- Some outliers due to missing data

  Bonn - Hb
Distributed Correlation

Comparison of DiFX output files main vs. branch correlator using diffDiFX.py

- Bonn - Ww
- Bonn - On
- Bonn - Sh
- Bonn - Hb
- Bonn - Vien

Difference on average ≤ 0.05%

Some outliers due to missing data
Distributed Correlation - Summary

• Issues (hiccups):
  • Wrong schedule for Onsala station: 256-16-1 instead of 512-16-2 - original database produced with 1-bit sampling.
  • One branch correlator used DiFX 2.5.1 instead of DiFX 2.5.2 (no tragedy ;-) ).
  • Two didn't apply the HOPS station codes table for difx2mark4; one forgot to use the control file.
  • Error during correlation: filelist contained two stations - needed recorrelation
  • Incomplete scans after e-transfer
  • Scans/baselines not correlated
  • Analysis of the VGOS database still pending
  • The test has confirmed that the results at the main and branch correlators are identical as expected.
  • Similarly, the analysis results should also be the same.

→ First attempt of DC (R1785) failed for above mentioned reasons (particularly number of scans differed) and beyond (e.g. test DiFX version was not considered).
VGOS today – state of the art

- Current status:
  - Antennas: GGAO, Westford, Kokee, Onsala, Wettzell, Yebes, (Ishioka)
  - Frequency range: 3 – 10 GHz, four bands
  - Dual-linear-polarization
  - Recording rate 8 Gbps
  - 30 sec scans (~50 scans per hour)
  - IVS VGOS tests: 24-hour observations with all available stations, correlated in Haystack
  - EU-VGOS tests: European stations, 4-hour observations, correlated in Bonn (main purposes: get to know backends and related issues)

- Achieved accuracy: WRMS deviation of the baseline length residuals about the weighted mean of 1.6 mm for baseline GGAO - Westford (Niell et al. 2018)
VGOS today – state of the art

#### VGOS Data Transmission and Correlation Plan (Petrachenko et al. 2015)

<table>
<thead>
<tr>
<th>Year</th>
<th># of sites</th>
<th>Hours of obs/day</th>
<th>data/day/site (TB)</th>
<th>data/day at correlator (TB)</th>
<th>network data rate at each site (Gbps)</th>
<th>network data rate at correlator (Gbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018</td>
<td>20</td>
<td>10</td>
<td>18.0</td>
<td>360</td>
<td>2.4</td>
<td>48</td>
</tr>
<tr>
<td>2019</td>
<td>24</td>
<td>12</td>
<td>21.6</td>
<td>518</td>
<td>2.8</td>
<td>68</td>
</tr>
<tr>
<td>2020</td>
<td>24</td>
<td>24</td>
<td>43.2</td>
<td>1037</td>
<td>5.6</td>
<td>134</td>
</tr>
</tbody>
</table>
## VGOS today - state of the art

### VGOS Data Transmission and Correlation Plan (Petrachenko et al. 2015)

<table>
<thead>
<tr>
<th>Year</th>
<th># of sites</th>
<th>Hours of obs/day</th>
<th>data/day/site (TB)</th>
<th>data/day at correlator (TB)</th>
<th>network data rate at each site (Gbps)</th>
<th>network data rate at correlator (Gbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018</td>
<td>20</td>
<td>10</td>
<td>18.0</td>
<td>360</td>
<td>2.4</td>
<td>48</td>
</tr>
<tr>
<td>2019</td>
<td>24</td>
<td>12</td>
<td>21.6</td>
<td>518</td>
<td>2.8</td>
<td>68</td>
</tr>
<tr>
<td>2020</td>
<td>24</td>
<td>24</td>
<td>43.2</td>
<td>1037</td>
<td>5.6</td>
<td>134</td>
</tr>
</tbody>
</table>

**Status today (2018):**

- # of sites 6; transfer rates 0.1 (Kk), 1 (GGAO, Wz), 10 (On, Ys), 20 (Wf) Gbps
- Hours of obs/day: bi-weekly 24-hour observations - data/site ~ 36 TB
- Network data rate at correlators:
  - Haystack 20 Gbps
  - Bonn 2 X 1 Gbps
  - WACO 1 Gbps - upgrade to 4 - 10 Gbps (not yet clear)
  - Shanghai 1 Gbps
  - Tsukuba 10 Gbps (non-DiFX)
### VGOS Data Transmission and Correlation Plan (Petrachenko et al. 2015)

<table>
<thead>
<tr>
<th>Year</th>
<th># of sites</th>
<th>Hours of obs/day</th>
<th>data/day/site (TB)</th>
<th>data/day at correlator (TB)</th>
<th>network data rate at each site (Gbps)</th>
<th>network data rate at correlator (Gbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018</td>
<td>20</td>
<td>10</td>
<td>18.0</td>
<td>360</td>
<td>2.4</td>
<td>48</td>
</tr>
<tr>
<td>2019</td>
<td>24</td>
<td>12</td>
<td>21.6</td>
<td>518</td>
<td>2.8</td>
<td>68</td>
</tr>
<tr>
<td>2020</td>
<td>24</td>
<td>24</td>
<td>43.2</td>
<td>1037</td>
<td>5.6</td>
<td>134</td>
</tr>
</tbody>
</table>

### Status today (2018):
- # of sites 6; transfer rates 0.1 (Kk), 1 (GGAO, Wz), 10 (On, Ys), 20 (Wf) Gbps
- Hours of obs/day: bi-weekly 24-hour observations - data/site ~ 36 TB
- Network data rate at correlators:
  - Haystack 20 Gbps
  - Bonn 2 X 1 Gbps
  - WACO 1 Gbps - upgrade to 4 - 10 Gbps (not yet clear)
  - Shanghai 1 Gbps
  - Tsukuba 10 Gbps (non-DiFX)

- IVS VGOS observations:
  - Moduls are shipped, takes 2 to 3 weeks, e-transfer On
  - Recording at 8 Gbps requires two Mark6 modules
- EU-VGOS tests:
  - Data are e-transferred, takes ~2 days/station (at 400 Mbps)
VGOS today - state of the art

- What amount of data (in terms of time) would be needed for a proper analysis, i.e. how many hours of observational data would be the minimum (2, 4, 6, 10, 12, 24 hours)?

- Analyze short sessions, ~1 hr, for UT1

- Normally for geodesy sessions a full 24 hours so that any diurnal effects will average out

- For VGOS data may be able to solve for piecewise continuous EOP's at perhaps 1 or 2 hour intervals, like done with clocks and atmospheres

- Probably at least 6 hours would be desireable

- **But:** analysts don't know the answer at this point
## Pros and Cons

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Pros and Cons

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sessions can be divided in proportion to the available capacity at each correlator</td>
<td></td>
</tr>
</tbody>
</table>
Pros and Cons

Pros
• Sessions can be divided in proportion to the available capacity at each correlator

Cons
• More complex logistics – stations must transfer/ship the raw data to various correlators which afterwards must upload correlation results to main correlator
Pros and Cons

Pros

- Sessions can be divided in proportion to the available capacity at each correlator

Cons

- More complex logistics - stations must transfer/ship the raw data to various correlators which afterwards must upload correlation results to main correlator
- If data are distributed in chunks of, e.g. 6 hours, main correlator can not easily correct for clock jumps - possibly need to fetch more data in the middle based on log information or on station's start/stop message - or in the end miss the clock jump due to incomplete information → need for recorrelation.
Pros and Cons

Pros

- Sessions can be divided in proportion to the available capacity at each correlator

- Astronomers already make use of DC, for example in case of EHT observations - divided into polarizations.

Cons

- More complex logistics - stations must transfer/ship the raw data to various correlators which afterwards must upload correlation results to main correlator

- If data are distributed in chunks of, e.g. 6 hours, main correlator can not easily correct for clock jumps - possibly need to fetch more data in the middle based on log information or on station's start/stop message - or in the end miss the clock jump due to incomplete information → need for recorrelation.
Pros and Cons

Pros

- Sessions can be divided in proportion to the available capacity at each correlator
- Astronomers already make use of DC, for example in case of EHT observations – divided into polarizations.

Cons

- More complex logistics - stations must transfer/ship the raw data to various correlators which afterwards must upload correlation results to main correlator
- If data are distributed in chunks of, e.g. 6 hours, main correlator can not easily correct for clock jumps - possibly need to fetch more data in the middle based on log information or on station's start/stop message - or in the end miss the clock jump due to incomplete information → need for recorrelation.
- The way, the VGOS data are recorded/written, doesn't allow for frequency separation.
Pros

• Sessions can be divided in proportion to the available capacity at each correlator

• Astronomers already make use of DC, for example in case of EHT observations - divided into polarizations.

Cons

• More complex logistics - stations must transfer/ship the raw data to various correlators which afterwards must upload correlation results to main correlator

• If data are distributed in chunks of, e.g. 6 hours, main correlator can not easily correct for clock jumps - possibly need to fetch more data in the middle based on log information or on station's start/stop message - or in the end miss the clock jump due to incomplete information → need for recorrelation.

• The way, the VGOS data are recorded/written, doesn't allow for frequency separation.

• Recording rate of 8 Gbps requires 2 Mark6 modules: could be filled up with several sessions, but prolongues turnaround time → desirable?
Pros and Cons

Pros

- Sessions can be divided in proportion to the available capacity at each correlator
- Astronomers already make use of DC, for example in case of EHT observations - divided into polarizations.

Cons

- More complex logistics - stations must transfer/ship the raw data to various correlators which afterwards must upload correlation results to main correlator
- If data are distributed in chunks of, e.g. 6 hours, main correlator can not easily correct for clock jumps - possibly need to fetch more data in the middle based on log information or on station's start/stop message - or in the end miss the clock jump due to incomplete information → need for recorrelation.
- The way, the VGOS data are recorded/written, doesn't allow for frequency separation.
- Recording rate of 8 Gbps requires 2 Mark6 modules: could be filled up with several sessions, but prolongues turnaround time → desirable? On the other hand, more shipments with half-empty disks is uneconomic. A possible solution, but also time-consuming: copy data from Mark6 onto flexbuff/raid and back onto module.
Pros and Cons

Pros

• Sessions can be divided in proportion to the available capacity at each correlator

• Astronomers already make use of DC, for example in case of EHT observations – divided into polarizations.

Cons

• More complex logistics - stations must transfer/ship the raw data to various correlators which afterwards must upload correlation results to main correlator

• If data are distributed in chunks of, e.g. 6 hours, main correlator can not easily correct for clock jumps – possibly need to fetch more data in the middle based on log information or on station’s start/stop message - or in the end miss the clock jump due to incomplete information → need for recorrelation.

• The way, the VGOS data are recorded/written, doesn't allow for frequency separation.

• Recording rate of 8 Gbps requires 2 Mark6 modules: could be filled up with several sessions, but prolongs turnaround time → desirable?

  On the other hand, more shipments with half-empty disks is uneconomic. A possible solution, but also time-consuming: copy data from Mark6 onto flexbuff/raid and back onto module.

• The necessary transfer rates for such huge amount of data cannot be met today or in the medium/long term because these are too expensive - e-transfer no viable solution
Conclusions and Prospects

- Does Distributed Correlation work?
  - In principle "Yes"
  - A distributed model will require very good project management and communication. The admin overhead will be higher than having a single correlator.
- How should the sessions be separated/distributed?
  - Chunks of, e.g., 6 hours:
    - Either four independently working branch correlators (might result in different clock values and control files for post-processing)
    - Or one main correlator responsible for fringe search and post-processing (requires more logistics, longer turnaround-time, issues during fringe search, e.g., with clock breaks)
  - Each 24-hour session sent to one correlator (our favourable suggestion in terms of shipment, station clocks/jumps, post-processing, maybe even turnaround time?)
  - Requires preferably seven correlators

→ Await feedback from the analysts
Conclusions and Prospects

● Does Distributed Correlation work?
  ● In principle "Yes"
  ● A distributed model will require very good project management and communication. The admin overhead will be higher than having a single correlator.
● How should the sessions be separated/distributed?
  ● Chunks of, e.g., 6 hours:
    ● Either four independently working branch correlators (might result in different clock values and control files for post-processing)
    ● Or one main correlator responsible for fringe search and post-processing (requires more logistics, longer turnaround-time, issues during fringe search, e.g., with clock breaks)
  ● Each 24-hour session sent to one correlator (our favourable suggestion in terms of shipment, station clocks/jumps, post-processing, maybe even turnaround time?)
  ● Requires preferably seven correlators

→ Thank you!
Cloud Computing

- Two main problems (Helge Rottmann, priv. comm.):
  - Bandwidth to transfer the data into the cloud
  - Costs of data storage in the cloud:
    - \(~ 2 - 4 \text{ cent/month} \rightarrow 1 \text{ TB} \sim 20 - 40 \text{ $/month} \rightarrow \text{okay for cm-VLBI}\)
    - But: expensive for broadband data (EHT, VGOS)
      - Example: EHT has \sim 7 \text{ PB per session} \rightarrow \sim 140.000 \text{ $/month, besides the transfer would take ages}\)

(Check e.g. https://aws.amazon.com/de/govcloud-us/pricing/s3/ for prices)