A GPU spectrometer for Radio Telescopes

Jongsoo Kim
Korea Astronomy and Space Science Institute and University of Science and Technology
EA-ALMA Development projects in KASI

- GPU Spectrometer for the ALMA TP Array
- Multi-beam receiver for the ALMA TP Array and ASTE
Fourier Transform
2.3. The Rise of Polyphase Filterbanks

A polyphase filterbank (PFB) is a computationally efficient structure used to create a band of digital filters. This is constructed from a polyphase finite impulse response (FIR) filter frontend that precedes an FFT; it offers far better isolation between channels than both XF and FX implementations (Figure 1). PFB-based correlators can still be considered ‘FX’ as the FFT can itself be considered a critically-sampled filterbank. The PFB architecture was first introduced by Schafer & Rabiner (1973), and further popularized by Bellanger et al. (1976).
Three Technologies for Correlator

• ASIC (Application-Specific Integrated Circuit)
  • Example, ALMA 64-antenna correlator

• FPGA (Field-Programmable Gate Arrays)
  • Example, ALMA ACA correlator

• Software (high level-languages, e.g., C/C++, MPI, CUDA/OpenCL)
  • Example, ALMA ACA spectrometer
Largest Correlators in the world

Table 1. Specifications of selected correlators with large figures of merit, \(N_{\text{ant}}^2B\) and \(N_{\text{ant}}B\).

<table>
<thead>
<tr>
<th>Telescope</th>
<th>Reference</th>
<th>(N_{\text{ant}})</th>
<th>(B) (GHz)</th>
<th>(N_{\text{ant}}B) (GHz)</th>
<th>(N_{\text{ant}}^2B) (GHz)</th>
<th>Data rate</th>
<th># of calculations</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHIME-1024</td>
<td>Vanderlinde et al. (2014)</td>
<td>1024</td>
<td>0.4</td>
<td>409.6</td>
<td>419430</td>
<td>FPGA+GPU</td>
<td></td>
</tr>
<tr>
<td>ALMA</td>
<td>Baudry et al. (2012)</td>
<td>64</td>
<td>16.0</td>
<td>1024.0</td>
<td>65536</td>
<td>ASIC</td>
<td></td>
</tr>
<tr>
<td>HERA-352</td>
<td>DeBoer et al. (in press)</td>
<td>352</td>
<td>0.2</td>
<td>70.4</td>
<td>24781</td>
<td>FPGA (CASPER)</td>
<td></td>
</tr>
<tr>
<td>ASKAP</td>
<td>Tuthill et al. (2014)</td>
<td>36</td>
<td>0.3</td>
<td>388.8(^\dagger)</td>
<td>13997(^\ddagger)</td>
<td>FPGA+GPU</td>
<td></td>
</tr>
<tr>
<td>eVLA</td>
<td>Perley et al. (2009)</td>
<td>27</td>
<td>8.0</td>
<td>216.0</td>
<td>5832</td>
<td>FPGA+GPU</td>
<td></td>
</tr>
<tr>
<td>LEDA</td>
<td>Kocz et al. (2015)</td>
<td>256</td>
<td>0.058</td>
<td>14.85</td>
<td>3801</td>
<td>FPGA+GPU</td>
<td></td>
</tr>
<tr>
<td>MeerKAT</td>
<td>Jonas (2009)</td>
<td>64</td>
<td>0.856</td>
<td>54.78</td>
<td>3506</td>
<td>FPGA+GPU</td>
<td></td>
</tr>
<tr>
<td>AARTFAAC-12</td>
<td>Prasad et al. (this issue)</td>
<td>576</td>
<td>6.25</td>
<td>3.6</td>
<td>2074</td>
<td>GPU</td>
<td></td>
</tr>
<tr>
<td>PAPER-128</td>
<td>Cheng et al. (2016)</td>
<td>128</td>
<td>0.100</td>
<td>12.8</td>
<td>1638</td>
<td>GPU</td>
<td></td>
</tr>
<tr>
<td>SMA</td>
<td>Primiani et al. (this issue)</td>
<td>8</td>
<td>16.0</td>
<td>128.0</td>
<td>1024</td>
<td>GPU</td>
<td></td>
</tr>
<tr>
<td>MWA</td>
<td>Ord et al. (2015)</td>
<td>128</td>
<td>0.030</td>
<td>3.84</td>
<td>492</td>
<td></td>
<td></td>
</tr>
<tr>
<td>uGMRT</td>
<td>Reddy et al. (submitted)</td>
<td>32</td>
<td>0.4</td>
<td>12.8</td>
<td>410</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EOVSA</td>
<td>Nita et al. (this issue)</td>
<td>16</td>
<td>0.6</td>
<td>96</td>
<td>154</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOFAR</td>
<td>de Vos et al. (2009)</td>
<td>48</td>
<td>0.032</td>
<td>1.54</td>
<td>74</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^\dagger\) Computed as \(N_{\text{beam}}N_{\text{ant}}B\), with the number of beams \(N_{\text{beam}}=36\).

\(^\ddagger\) Computed as \(N_{\text{beam}}N_{\text{ant}}^2B\) with \(N_{\text{beam}}=36\).
Pros/Cons of Software Correlator

• Pros of software spectrometer
  • rapid and easy development
  • flexibility (e.g., RFI) and expandability
  • 32bit floating point operations (high-precision)
  • Commodity Off-The-Shelf Technology (COTS)

• Disadvantages of software spectrometer
  • low performance/Watt
NVIDIA GPU Roadmap

Tesla V100:
single precision 14 Tflops

Tesla P100:
single precision 10.6 Tflops

Geforce Titan X:
single precision 6.6 Tflops

Tesla K40:
single precision 5.04 Tflops
Prototype GPU spectrometer
Data Flow in Spectrometer

Input data are in VDIF format or in a yet to be specified ACA TP DXRP PCIe format. Sources: network, file, on-the-fly synthetic generator.
GeForce Titan X
- Maxwell architecture
- CUDA cores: 3072
- Base and boost clocks: 1000, 1075 MHz
- Performance: 6.14~6.6 Tflops single precision
- Memory Bandwidth: 336.5 GB/sec
- Memory: 12GB

GeForce GTX 980
- Maxwell architecture
- CUDA cores: 2048
- Base clock: 1064 MHz
- Performance: 4.36 Tflops single precision
- Memory Bandwidth: 224 GB/sec
- Memory: 4 GB

Tesla K40m
- Kepler architecture
- Cores: 2880
- Base, boost clocks: 745 MHz, 810 MHz and 875 MHz
- Performance: 4.29~5.04 Tflops single precision
- Memory bandwidth: 288 GB/sec
- Memory: 12 GB
data copy from CPU to GPU

- converges to 12.5 GB/sec < 16GB/sec (PCIE 3)
- 2 bits/sample: 12GB/sec —> 48 G samples/sec (24 GHz bandwidth)
- 3 bits/sample: 12GB/sec —> 32 G samples/sec (16 GHz bandwidth)
- 4 bits/sample: 12GB/sec —> 24 G samples/sec (12 GHz bandwidth)
Data bit conversion (2 or 3 bits to 32bits)

- lookup table for a 2 bit
  - \{-3.3359, -1.0, +1.0, +3.3359\}
- lookup table for a 3 bit sample
  - \{-7.0f, -5.0f, -1.0f, -3.0f, +7.0f, +5.0f, +1.0f, +3.0f\}
- ~ 50 Gsamples/s for 2bit and 3bit samples with GTX Titan X
cuFFT

cufftPlanMany(....);
cufftExecR2C(plan, idata, odata)

• total number of samples: 250*2^20
• 7.5~17.5 Gsample/sec (3.8~8.8 GHz in single polarization)
• ACA correlation used 2^20 point FFT
• for a given number of samples, FFT performance is higher at small fft points
FFT performance of different GPU cards
Test results of using the NRO 45-m antenna

SiO (v=2, J=1-0) @ 42.8GHz

K-GPU Spectrometer

SiO (v=1, J=1-0) @ 43.1GHz

SAM45
Results of test observation using KVN Yonsei antenna

Orion KL H2O and SiO maser lines
Conclusions

• The performance of GPU is good enough to make a spectrometer for a single dish antenna or a correlator for an array with modest number of antennas.

• A GPU spectrometer (one server with four GPU cards) could process data streams of 32Gsamples/s from one ALMA antenna in real time.
  • The most time-consuming part is FFT, but cuFFT (in CUDA FFT library) is fast enough for our spectrometer.