Memo on COSMIC GPU cluster design

Cherry (Nov, 2021)

Min. No. of GPU nodes required

- Ethernet output data (Nyquist sampled) rate:
  2 GHz * 2 complex * 2 pol * 8 bits * 27 antennas = 1728 Gbps
- Each compute node with 2x 100 GbE interfaces operates in “ping pong mode”
  Data rate per compute node: 1728 / 100 Gbps so need >17 nodes , or conservatively assume 75 Gbps BW (based on MeerKAT benchmark from Dave) then >24 nodes
- Assuming 32 nodes (which is a nice power of 2 number), each node will have:
  - Input rate = 6.75 GB/s
  - Input rate per NIC (assuming 2 NICs per node) = 3.4 GB/s
  - NVMe for 5 min = 2 TB

GPU model

- RTX A4000 , 4 GPUs per node
- 19.2 Tflops FP32 per card = 76.8 Tflops per node

Flops calculation

Upchannelize

$5 * (N * \ln N) / N * df * n_{pol} * n_{element} * n_{chan}$

Beamforming

$8 * n_{beam} * n_{ants} * n_{pol} * df * n_{chan}$

Refer to memo MeerKAT DSP for more background information. As described in that document, the beamforming cost is constant irrespective of the frequency mode. It only scales with the number of coherent beams we will form and the no. of antennas to combine. Regarding the Upchannelization FFT cost, it is slightly less when there are more coarse channels to start with (shorter upchan FFT required). However, the difference is not much. Table 1 shows the Gflops required for the beamforming and upchannelization steps. Column 1 lists the case of MeerKAT L-band obs at 1k mode (the worst case scenario), which amounts to ~1000 Gflops and will take up ~<10% GPU resources on the Nvidia 2080 card that we have at MeerKAT.

For COSMIC, the current plan is to form 64 coherent beams. In Table 1, three different COSMIC configurations are listed. Note that we will have more powerful GPUs for COSMIC. The current expectation is to get the RTX A4000 which has 19.2 Tflops per GPU, and we can fit four (4) cards per node, which gives us plenty of GPU compute. Using 32 GPU nodes (see column 2), we should be able to comfortably perform the beamforming and upchannelization with everything using only a few % of a GPU node.

	MeerKAT (op1)	COSMIC (32 nodes 64 beams)	COSMIC (64 nodes 64 beams)	COSMIC (24 nodes 64 beams)	COSMIC (32 nodes 27 beams)
nbeam	64	64	64	64	27
nant	64	27	27	27	27
npol	2	2	2	2	2
df (Hz)	1.594424248	1.6	1.6	1.6	1.6
total BW (MHz)	856	2048	2048	2048	2048
no. nodes	64	32	64	24	32
total nchan per node	8388608	40000000	20000000	53333333.33	40000000
BW per node (MHz)	13.375	64	32	85.33333333	64
Beamform (Gflops)	876.544	1769.472	884.736	2359.296	746.496
1k mode					
no. coarse chan total	1024	1024	1024	1024	1024
N	524288	1250000	1250000	1250000	1250000
log base 2	19	20.25349666	20.25349666	20.25349666	20.25349666
df (Hz) per coarse chan	8.36E+05	2.00E+06	2.00E+06	2.00E+06	2.00E+06
nchan coarse per node	16	32	16	42.66666667	32
Upchan (Gflops)	112.73	242.59	121.29	323.45	242.59
Total (Gflops)	**989.28**	2,012.06	**1,006.03**	**2,682.75**	**989.08**
	13.5	76.8	76.8	76.8	76.8
% GPU usage	7.156231608	2.558466521	1.27923326	3.411288695	1.257685271

Table 1) Gflops required for MeerKAT and COSMIC on beamforming and upchannelization.

GUPPI raw format

We will be using GUPPI raw format as input data for the GPU processing. Software will be required to convert the FPGA output packet into GUPPI raw format. The understanding is that such code exists for the ATA (written by Ross) and should be adaptable for COSMIC. Table 2 shows the GUPPI raw data block sizes for MeerkAT, ATA and the suggested COSMIC system. Note that we aim to create an effective block size that is O(128) MiB. With 32 coarse channels per node at the 32 node configuration, we can either fit 2^15 time samples per block summing to 17ms per block for a 108MiB block size, or fit 2^16 time samples per block summing to 34ms per block for a 216MiB block size (see last two columns in Table 2).

	MeerKAT 1k	MeerKAT 4k	MeerKAT 32k	ATA wide 4b	ATA wide 8b	ATA bf 4b	COSMIC (64 node, 1k mode)	COSMIC (32 node, 1k mode)	COSMIC (32 node, 1k mode)
N. bits	8	8	8	4	8	8	8	8	8
N. pol	2	2	2	2	2	2	2	2	2
N. time	32768	8192	1024	8192	4096	8192	65536	32768	65536
dt (us)	1.196	4.785	38.280				0.512	0.512	0.512
equivalent to time (s)	0.039	0.039	0.039				0.034	0.017	0.034
N. freq per node	16	64	512	1280	1280	192	16	32	32
N. ant	64	64	64	5	5	20	27	27	27
"Effective" Block size (MiB)	128	128	128	100	100	120	108	108	216

Table 2) GUPPI raw data block size comparison between MK, ATA, and COSMIC.

Executive summary

We target a 32-node GPU cluster, each node consists of 4 RTX A4000 GPU cards. We will form 64 coherent beams. The beamforming and upchannelization combined should in theory require only a few % of GPU usage.

The FPGA will produce 1024 coarse channels at a frequency resolution of 2 MHz per channel. The frequency resolution is kept constant no matter whether we are in 3-bit or 8-bit mode. Each GPU node will have 32 coarse channels. To achieve ~1Hz fine frequency resolution, we will likely need upchannelization FFT of N=2^20.

The GUPPI raw data block can have either 2^15 or 2^16 samples per block (TBD).