THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

FPGA-based Acceleration of SKA Pulsar Search Modules Using OpenCL

Haomiao Wang and Tyrone Sherwin

12/07/2018

PARALLEL AND RECONFIGURABLE COMPUTING LAB

Pulsar Search

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

Pre-processed Digital Data

Central Signal Processor (CSP)

- Correlator & Beamformer (CBF)
- Pulsar Search Engine (PSS)
- Pulsar Timing Engine (PST)

Science Data Processor (SDP)

SKA-Mid Africa Widefield Image credit: SKA Organisation

- Test of general relativity and gravitational waves => Pulsar searches
- Pulsar Search Engine (PSS)
  - Single Pulse Search
  - Time-domain acceleration search
  - Frequency-domain acceleration search

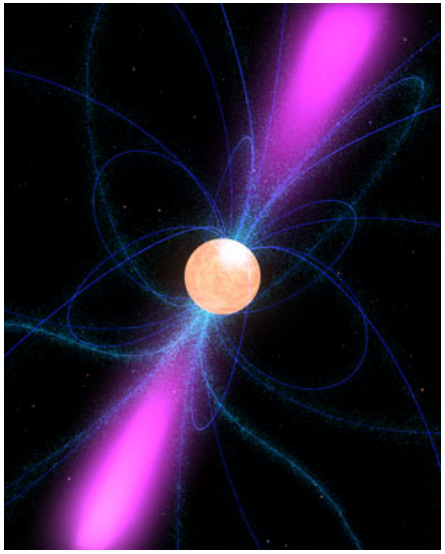
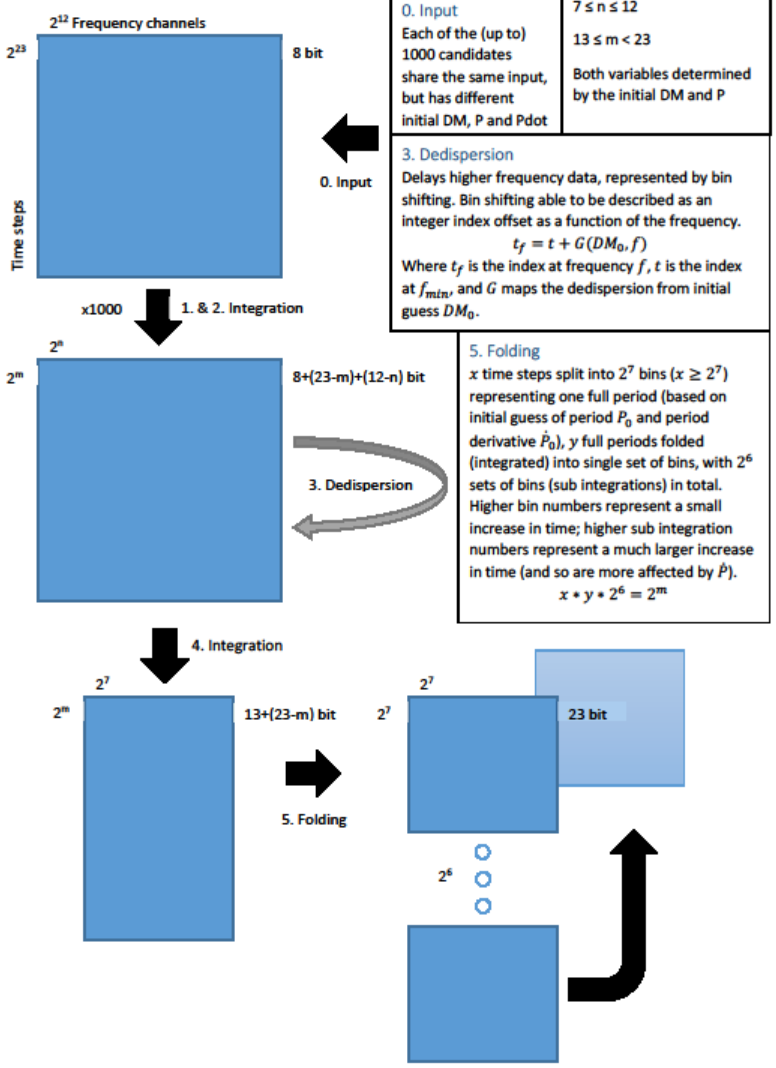


Image credit: NASA

Example-FLDO

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

1-5 Preprocessing (Dedispersion & Folding)



0. Input

Each of the (up to) 1000 candidates share the same input, but has different initial DM, P and Pdot

$7 \le n \le 12$

$13 \le m < 23$

Both variables determined by the initial DM and P

3. Dedispersion

Delays higher frequency data, represented by bin shifting. Bin shifting able to be described as an integer index offset as a function of the frequency.

$$t_f = t + G(DM_0, f)$$

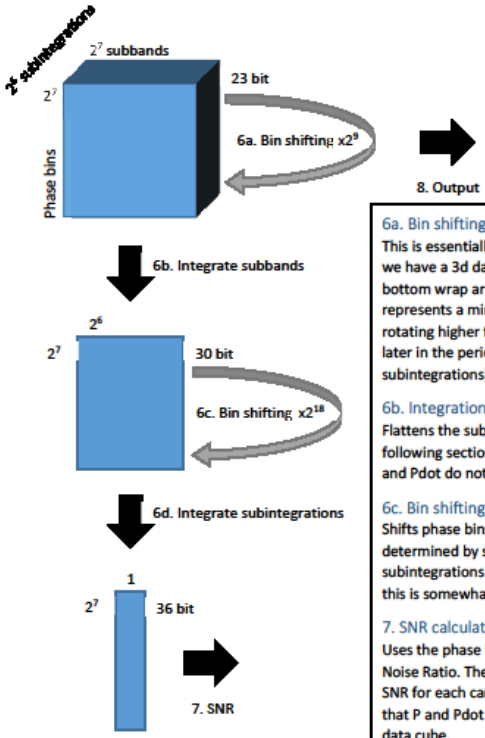
Where t_f is the index at frequency f, t is the index at f_{min}, and G maps the dedispersion from initial guess DM_0.

5. Folding

x time steps split into 2^7 bins ($x \geq 2^7$) representing one full period (based on initial guess of period P_0 and period derivative $\dot{P}_0$), y full periods folded (integrated) into single set of bins, with 2^6 sets of bins (sub integrations) in total. Higher bin numbers represent a small increase in time; higher sub integration numbers represent a much larger increase in time (and so are more affected by $\dot{P}$).

$$x * y * 2^6 = 2^m$$

6-7 Optimisation



6a. Bin shifting (DM)

This is essentially the same as step 3, except here we have a 3d data cube and the values at the bottom wrap around to be at the top. This represents a minor adjustment to the DM by rotating higher frequency values to bins which are later in the period. Shifts phase bins (across all subintegrations) as determined by the subband.

6b. Integration (subband)

Flattens the subband axis to reduce computation in following sections. This is possible as changes to P and Pdot do not affect this axis.

6c. Bin shifting (P & Pdot)

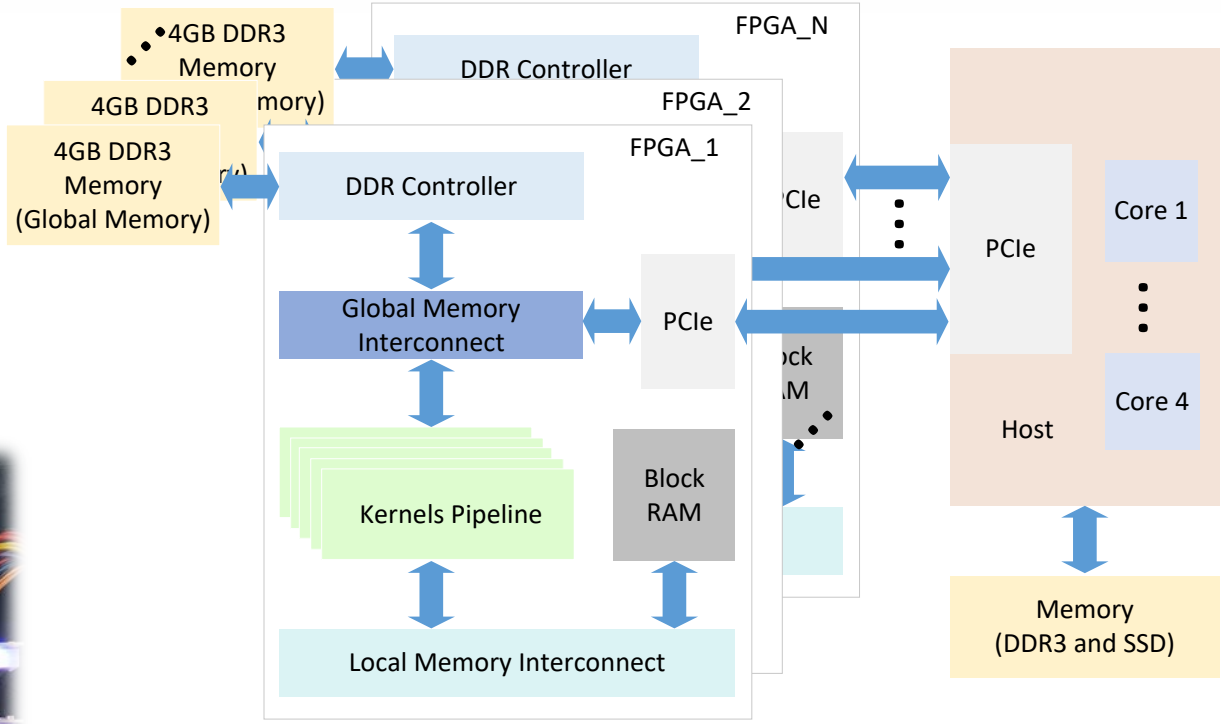
Shifts phase bins (across all subbands) as determined by subintegration. Since subintegrations are in effect a part of the time axis, this is somewhat equivalent to an acceleration.

7. SNR calculation (and 8. Output)

Uses the phase bin axis to calculate the Signal to Noise Ratio. The profile (data cube) with the best SNR for each candidate is then output, requiring that P and Pdot shifting be applied to the relevant data cube.

OpenCL for FPGA

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau



GeForce GTX TITAN X

Intel Stratix 10 FPGA

Overview

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

Designing pipeline stages of Pulsar Search Pipeline

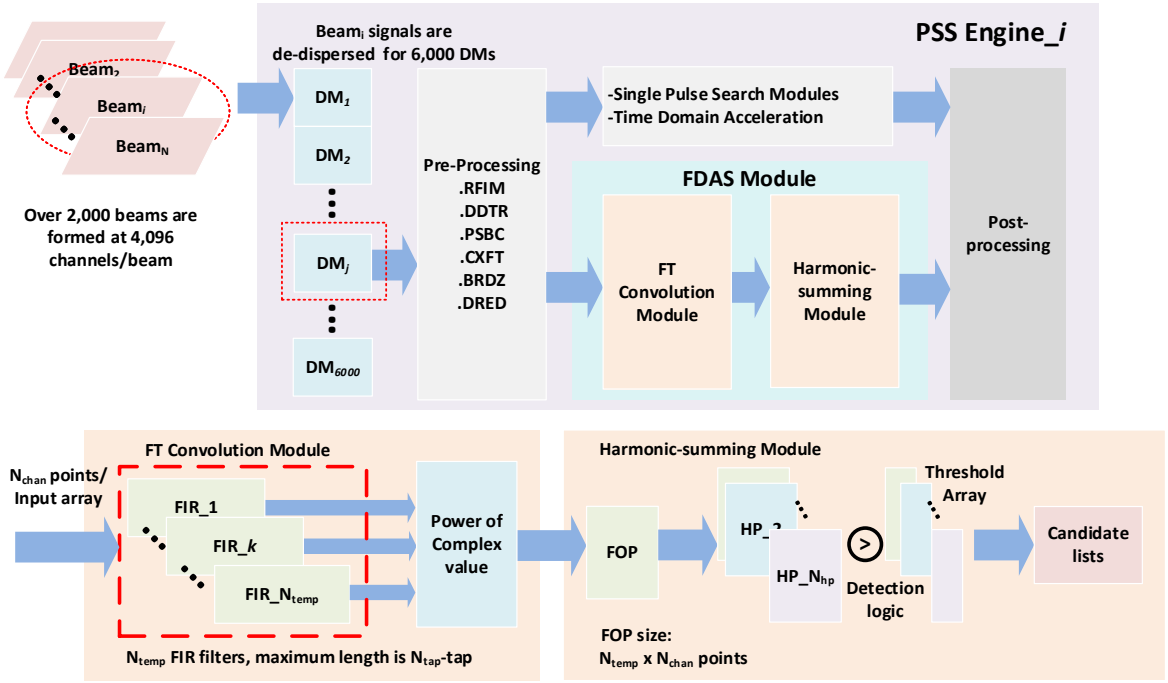
- Fourier Domain Acceleration Search (FDAS)
  - FT convolution
  - Harmonic summing
- RFI Mitigation (RFIM)
  - Median filtering

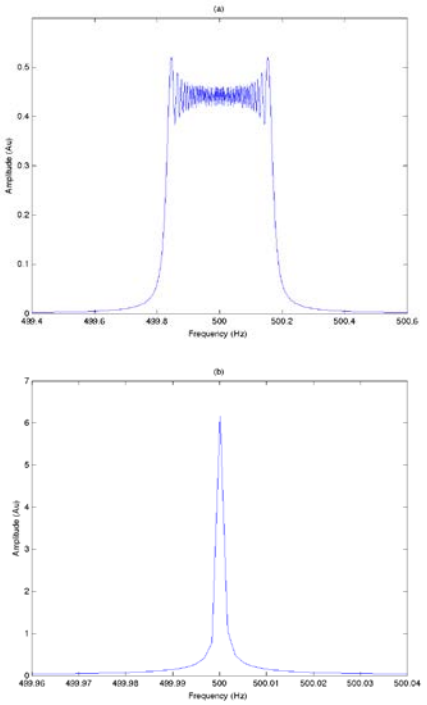
FDAS Module

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

- Search for (binary) pulsars with constant frequency derivatives
- Matched filtering (a group of FIR filters with different lengths, complex floating-point)

$$A_{r_0} = \sum_{k=[r_0]-\frac{m}{2}}^{[r_0]+\frac{m}{2}} A_k A^*_{r_0-k}$$



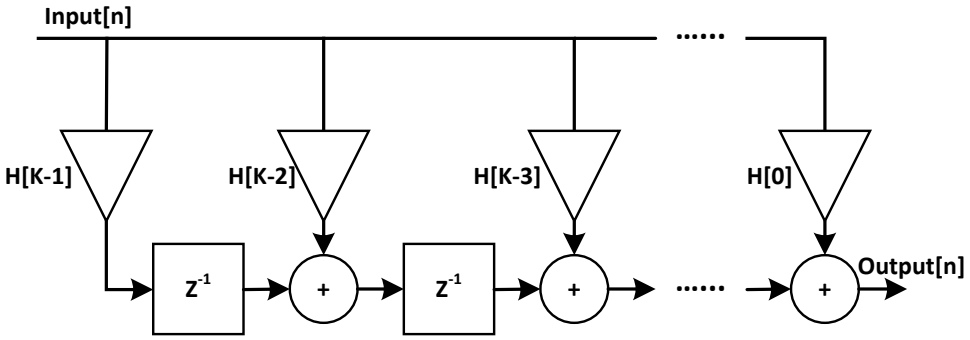


TD and FD FIR Filter

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

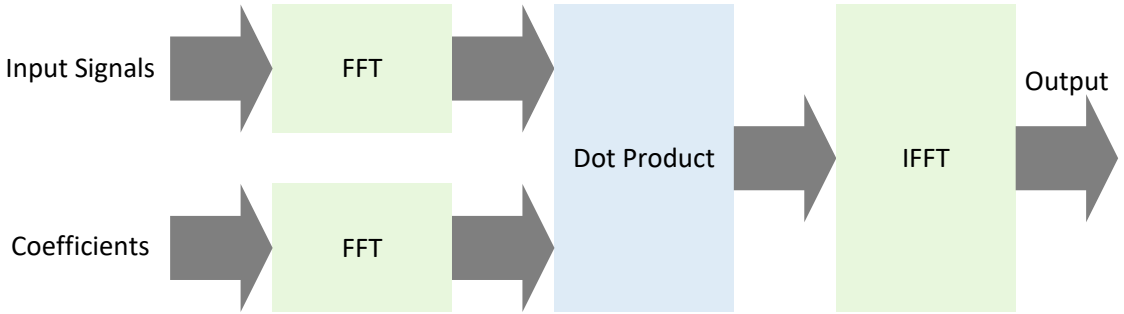
- Time-domain FIR Filter (**TDFIR**)

$$y_m[i] = \sum_{k=0}^{K-1} x_m[i-k]h_m[k], \qquad for\ i = 0,1,\dots N-1$$



- Frequency-domain FIR Filter (**FDFIR**)

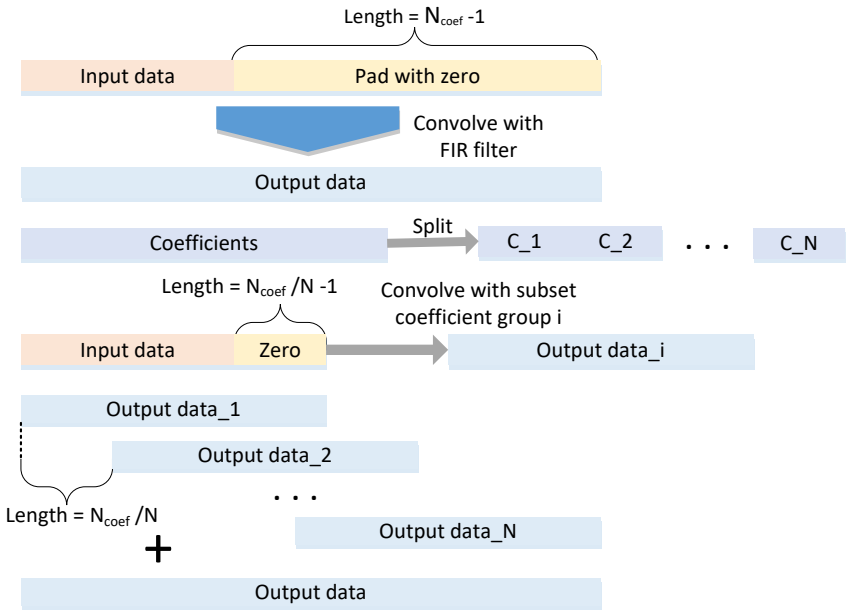
$$x * h = F^{-1}\{F\{x\} \cdot F\{h\}\}$$



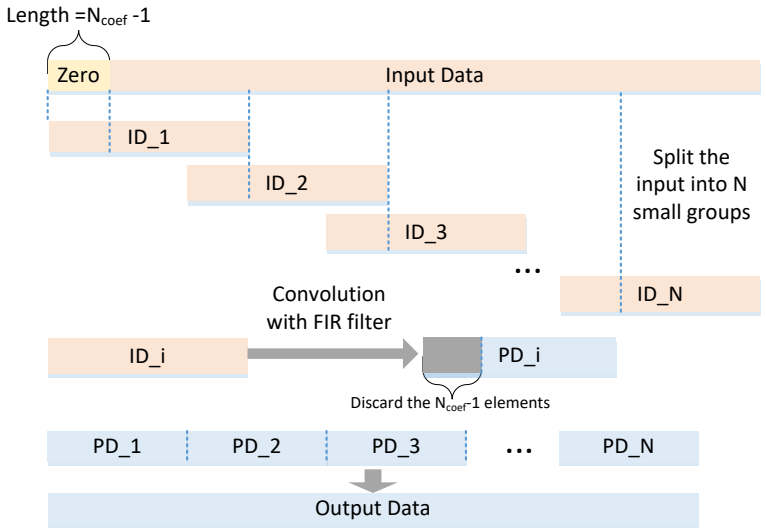
OLA and OLS Algorithms

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

- **Overlap-add algorithm (OLA)**
  - split the **coefficient array** (TDFIR)
- **Overlap-save algorithm (OLS)**
  - split the **input array** (FDFIR)



Overlap-add Algorithm

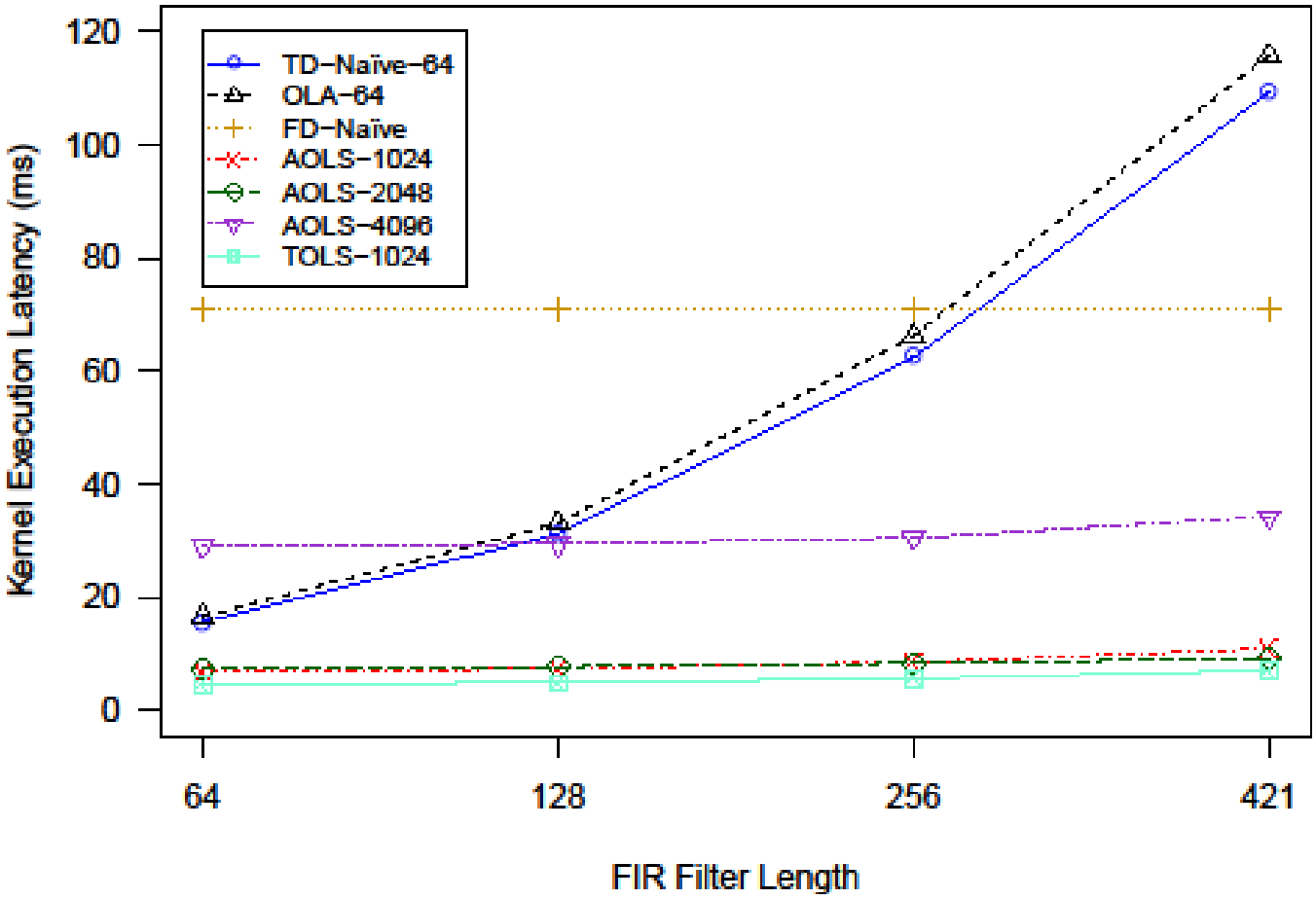


Overlap-save Algorithm

Latency-TDFIR vs. FDFIR

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

Latencies of a single FPGA (Intel Stratix V A7) in processing same input array using 7 different methods

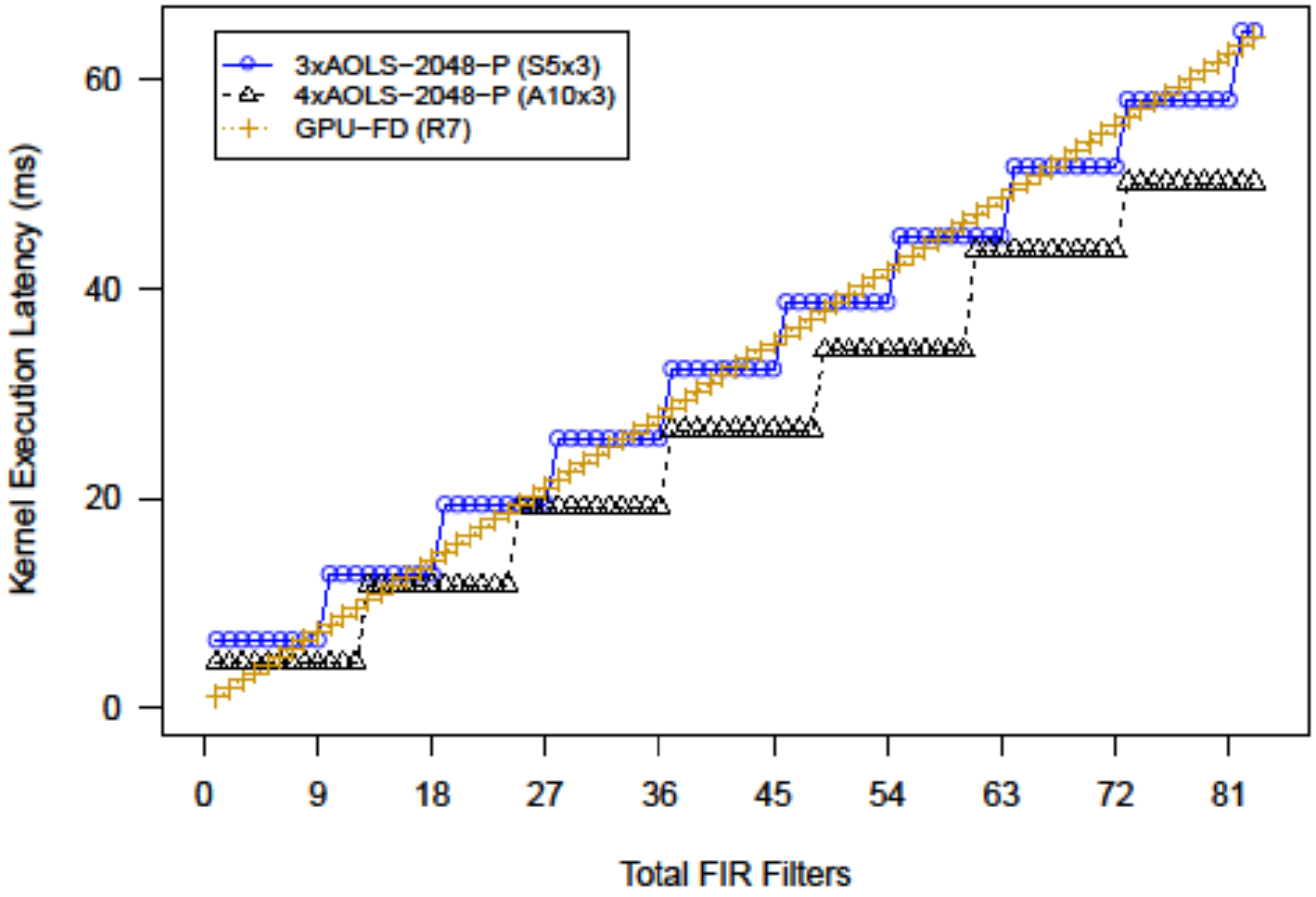


Latency-FPGA vs. GPU

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

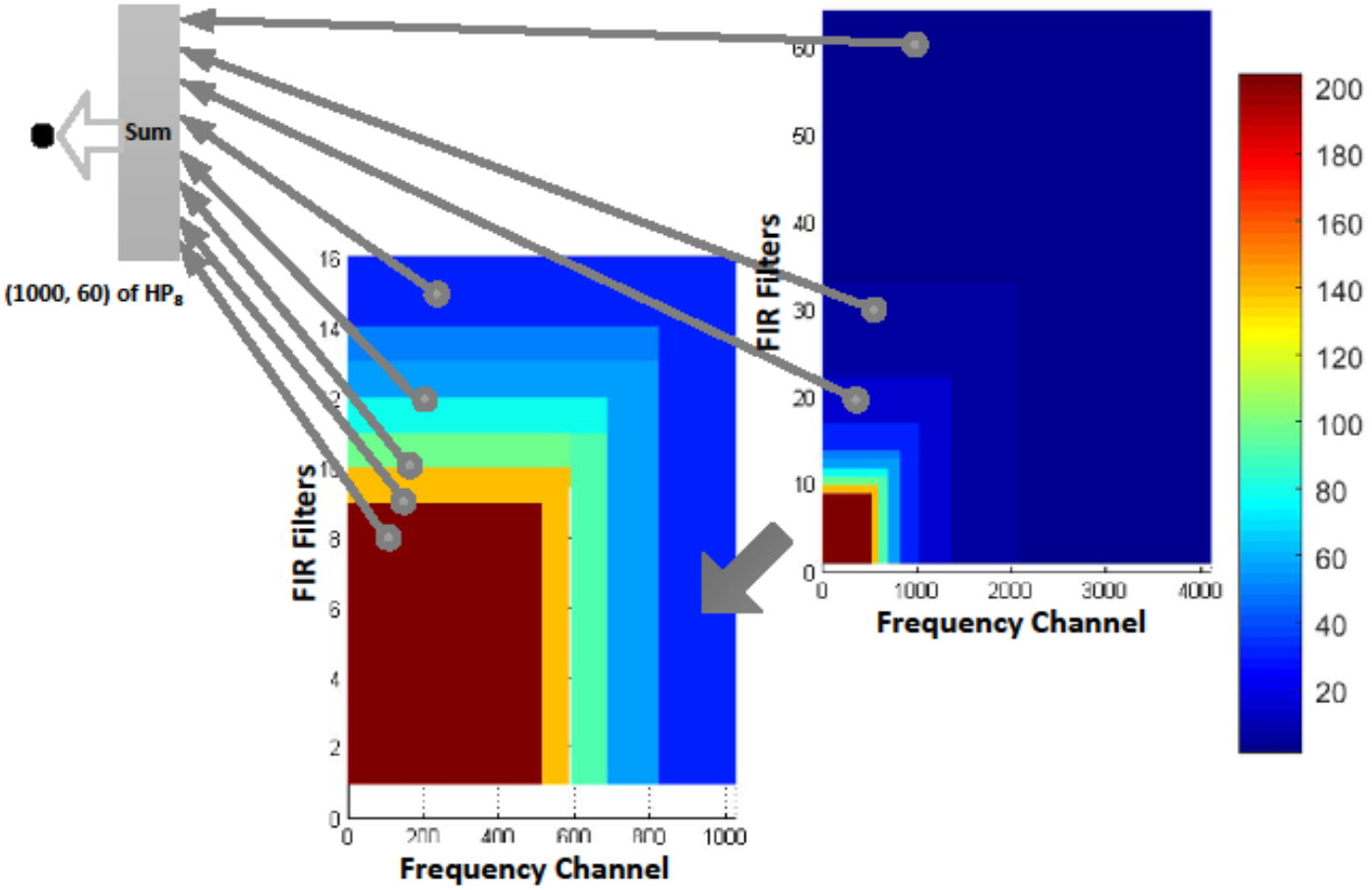
Latencies of a single GPU (AMD Radeon R7) and 3 FPGAs (Intel Stratix V and Arria 10)

3xFIRs on one Stratix V image and 4xFIRs on one Arria 10 image



THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

Optimizing irregular accesses to off-chip memory



Processing Methods

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

- **SingleHP method**

  Processing a single harmonic plane at a time

- **MultipleHP method**

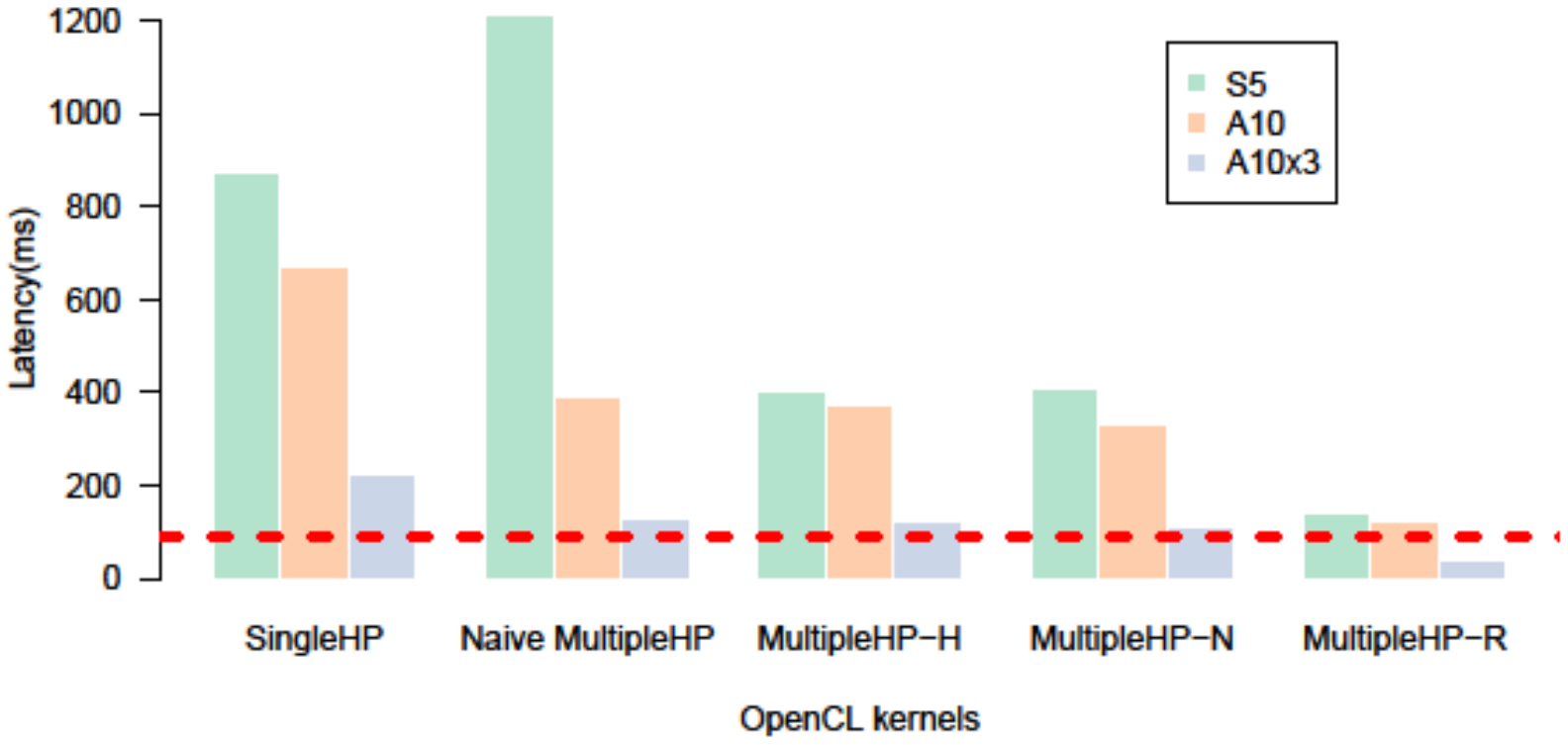
  Processing multiple harmonic planes in parallel

  - Naïve MultipleHP
  - **MultipleHP-H**: preloading data with high toughing frequency
  - **MultipleHP-N**: loading necessary data to calculate a block of points
  - **MultipleHP-R**: based on MultipleHP-N and the FOP is reordered to achieve stream mode

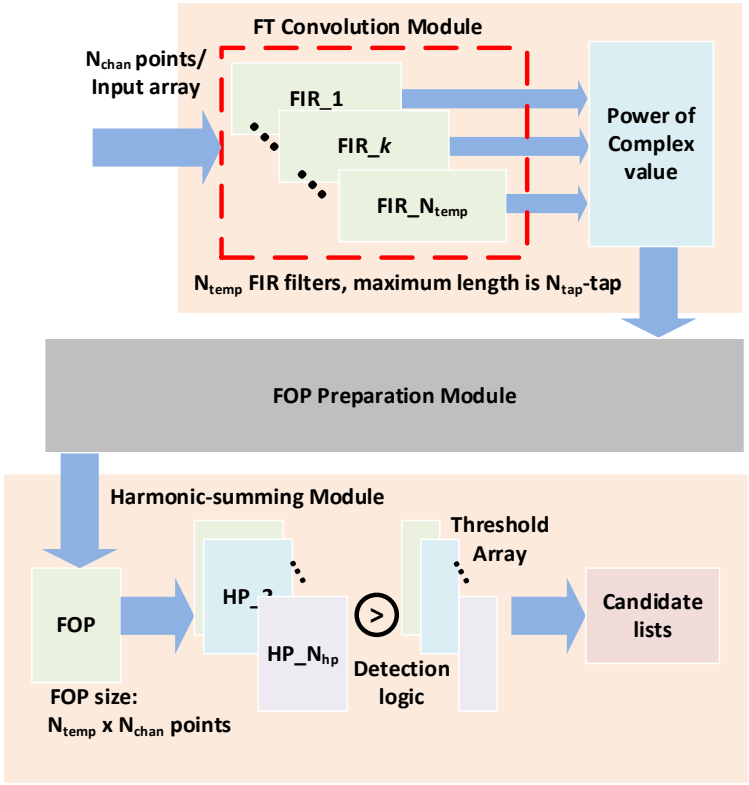
Execution Latency

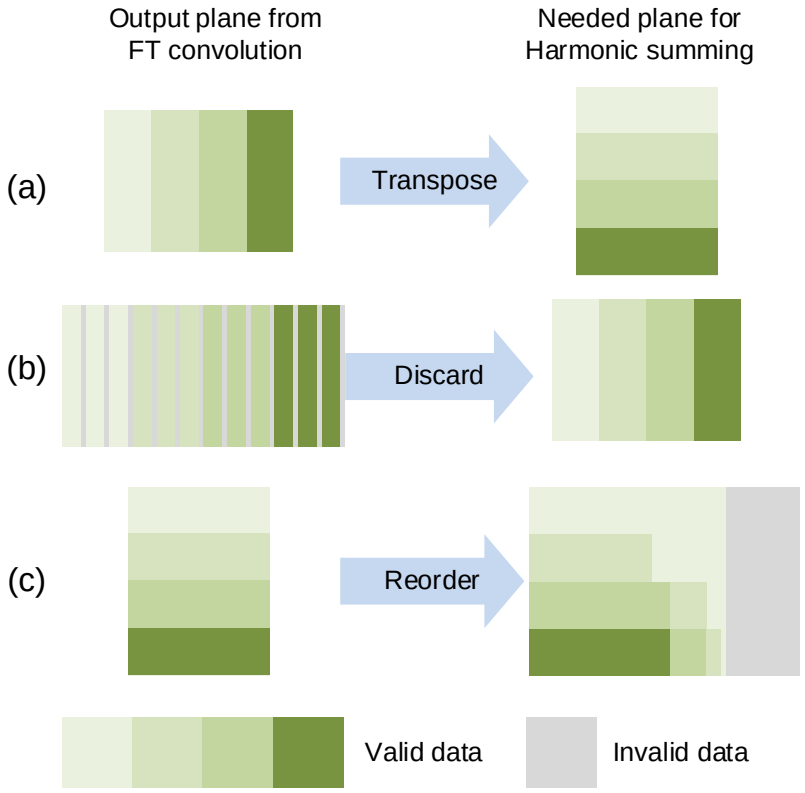
THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

MultipleHP-R method performs best among the proposed methods and it meets the required time limitation



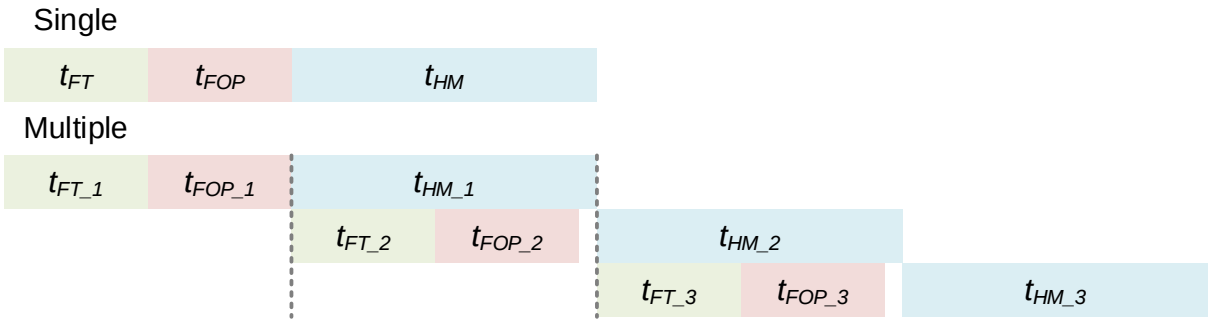
An FOP preparation module is added to connect the FT convolution module and harmonic summing module



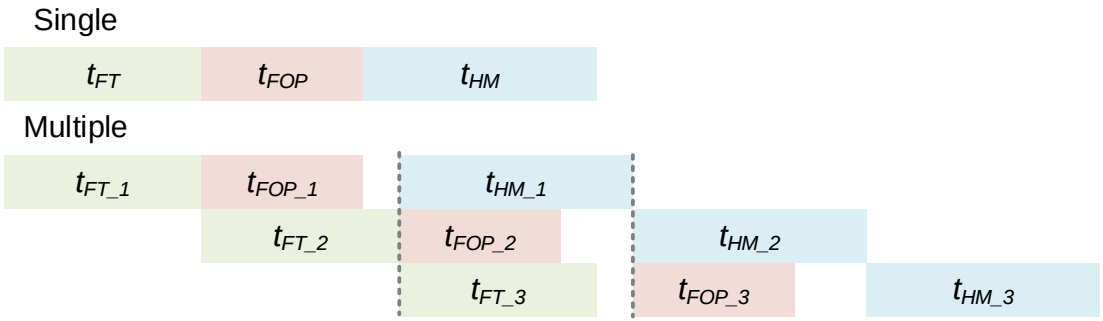


Pipeline Computing

$Max(t_{FT}, t_{FOP}, t_{HM}) \geq t_{FDAS}/2$



$Max(t_{FT}, t_{FOP}, t_{HM}) < t_{FDAS}/2$



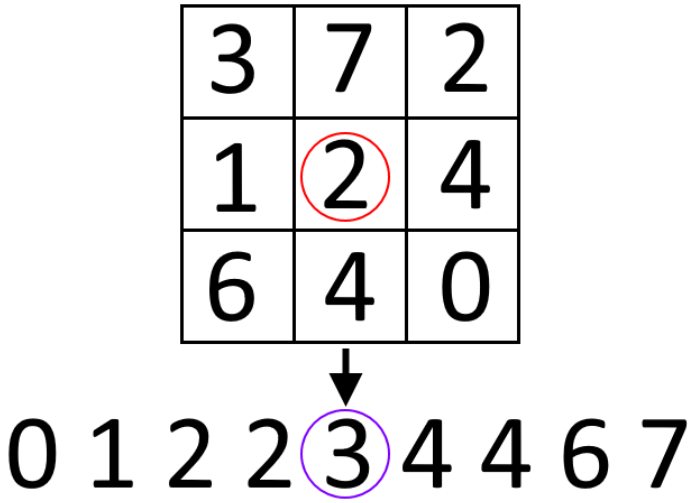
Median Filtering

Form of stencil computation:

- Recomputing an array element based on neighbour values
- In median filter, target value is replaced by median of all values in a window around target value
- => sorting all values, tacking middle value

Example:

- 2D array, window (stencil) is 3 x 3



Requirements-SKA Mid

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

Large window size.

- 63 frequency channels, 5MHz, *n*.
- 1023 time steps, 65ms, *m*.

Array Size.

- Frequency coverage 4096 channels, 300MHz.
- Time coverage potentially infinite, steps of 64*us*.

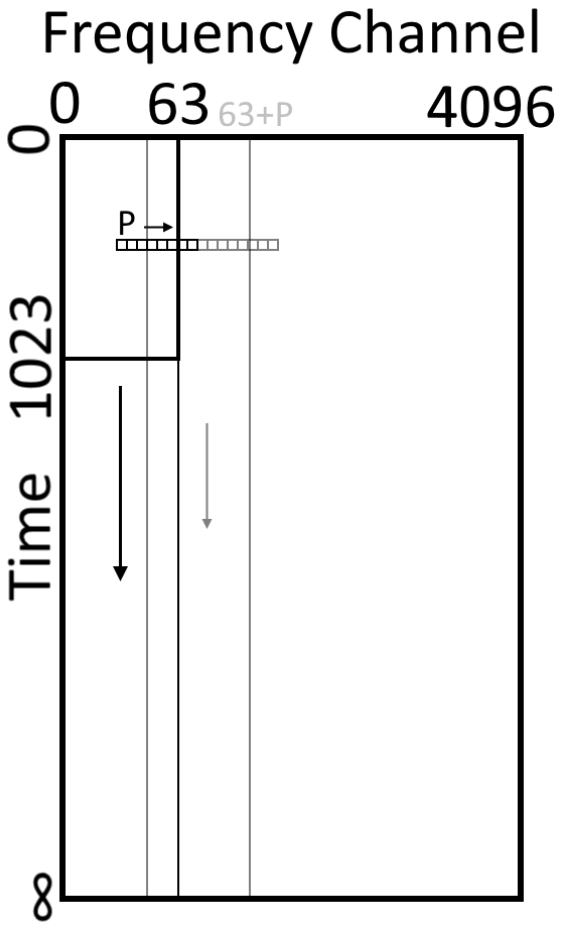
Throughput requirements.

- 8bit input data.
- 4096 channels every 64*us*.
- 64 Million values per second.

Algorithmic Approach

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

- Choice of algorithmic approach dominates performance
- Using best of various state-of-the-art methods
- Histograms
  - 256 bins for 8bit data
  - Each bin counts that value in current window.
- Sliding
  - Adjacent windows overlap in data
  - Reuse previous window, replacing oldest row with next row

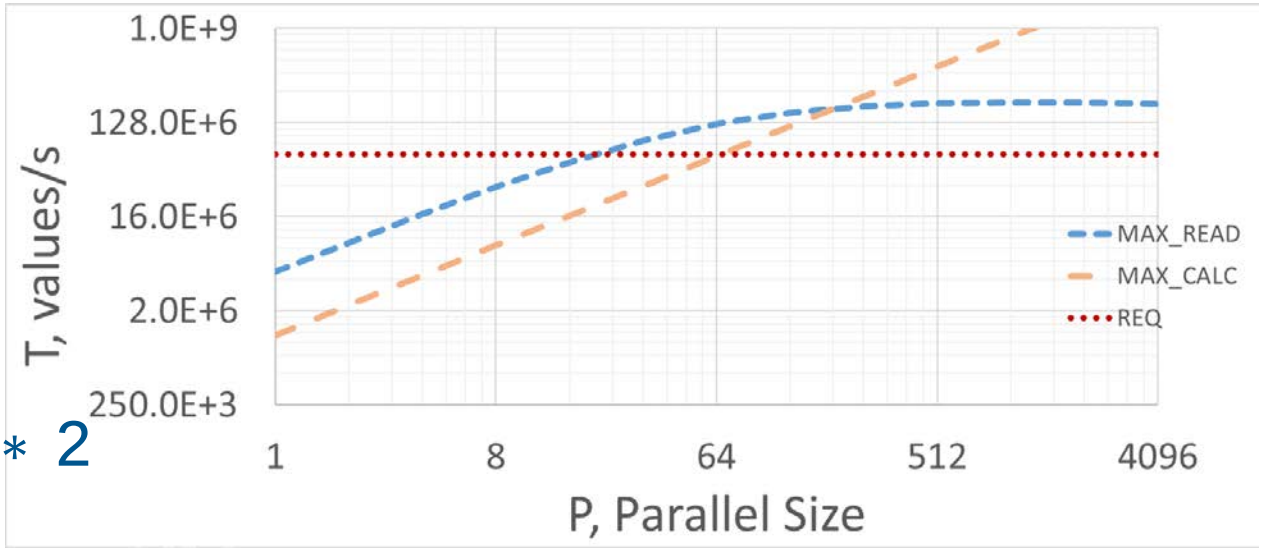


Roofline Model

- Developed to determine theoretical performance.
- Limitations placed on loads per cycle and calculations per cycle.
- $n=63$ requires $P>64$.

T is throughput, f is frequency, L is loads per row, I is loads per cycle

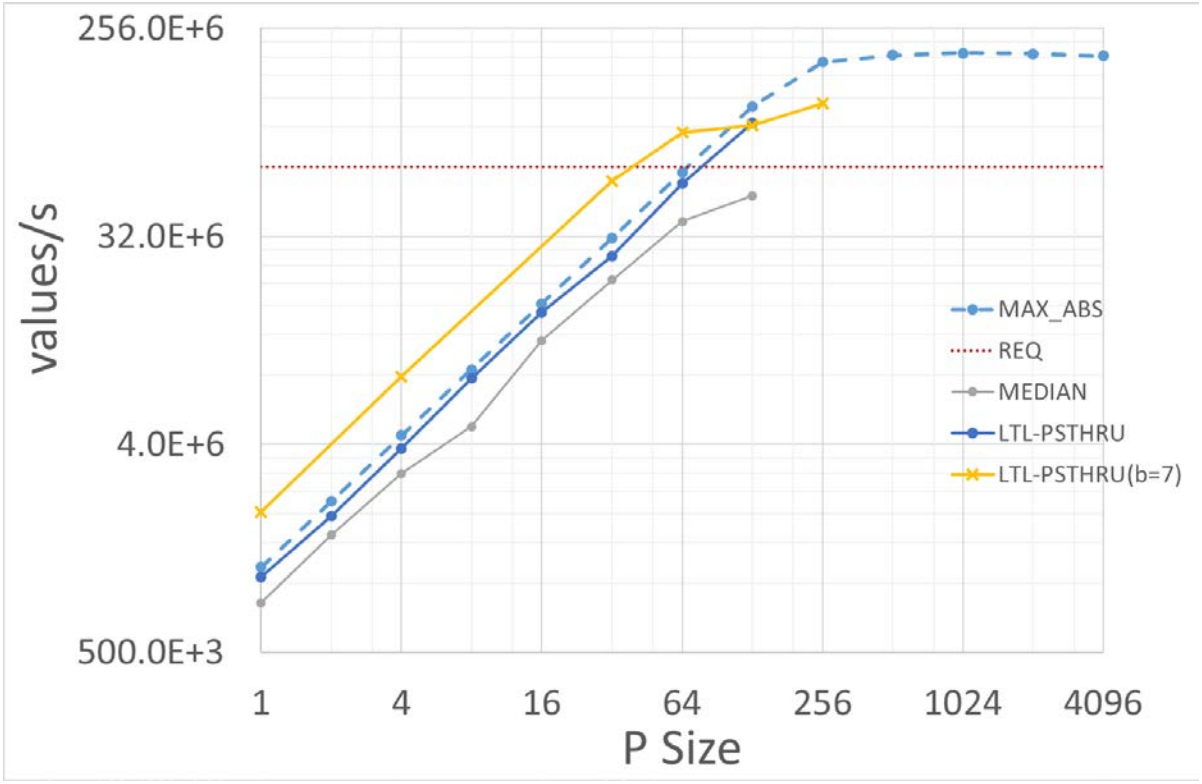
$$T = (V/c) * f$$
$$V/c = I/L * P$$
$$L = (n+P-1) * 2$$



FPGA Performance

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

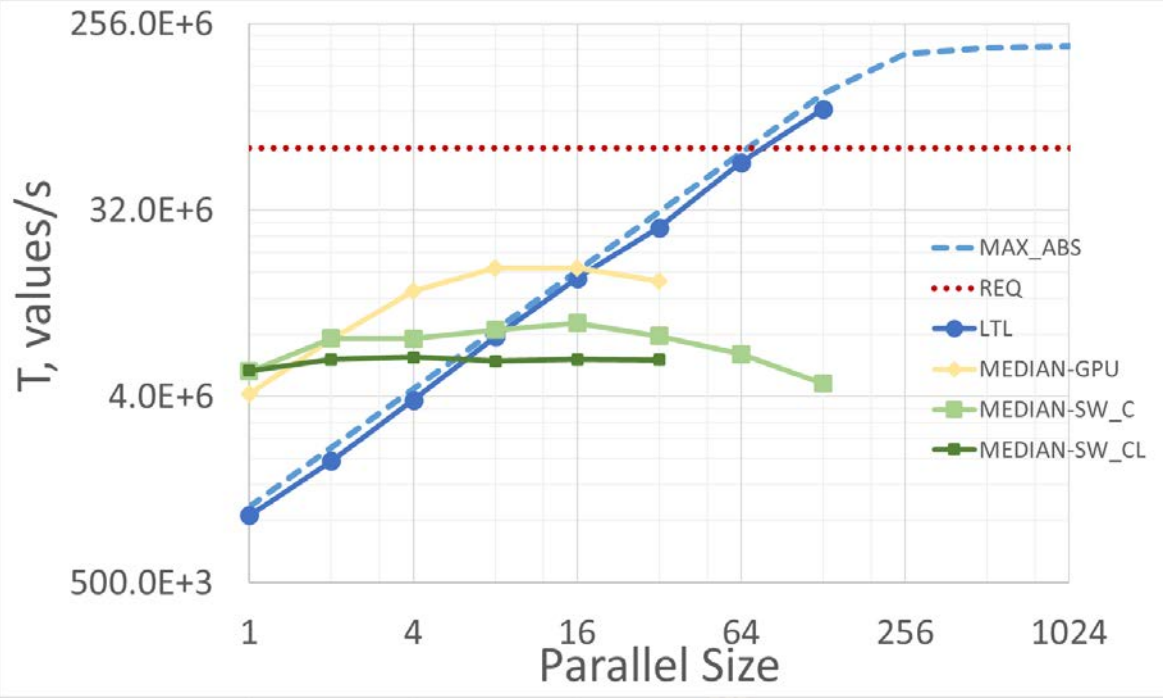
- Pipeline balancing
- Able to achieve required throughput
- LTL-PSTHRU(P=128) reaching 99.3Mvalue/s.
  - Theoretical performance of 117.7M, limited by calculation rate.
  - 84.4% → Room for improvement.
- LTL-PSTHRU(P=128 , b =7) reduces bit depth for higher performance



Comparison with GPU&CPU

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

- CPU & GPU versions not able to perform as well – despite OpenCL adjustments
- CPU appears not to operate in parallel
- GPU quickly drops off
  - Appears similar in shape to roofline models
  - Suggests possible resource constraints
- FPGA starts slower, much better scaling with *P*



Conclusions

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

- Investigated the optimization of FT convolution and harmonic-summing and median filtering modules
- Pipelining modules on single device (no host-FPGA communication)
- Using multiple FPGAs
- High-level approach is employed to implement the optimized methods
  - Covered large design space
  - Easy porting and sharing with partners

Future Tasks

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

- Designing FPGA implementation for FLDO – Candidate Folding and Optimisation

Long term:

- Creating OpenCL library (target FPGA) for all (computationally intensive) Pulsar Search Pipeline stages

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

Thanks

Q&A