Accelerating Compute with FPGAs

based Xilinx accelerator cards

Xinping Deng

10 September 2020, CASS Co-learnium

ASTRONOMY AND SPACE SCIENCE
www.csiro.au

Kind of introduction …

CPU, GPU and FPGA

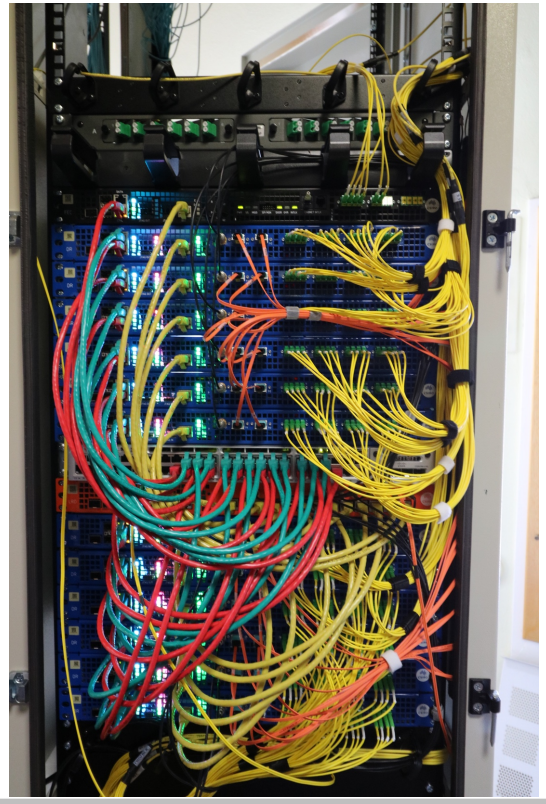
A **central processing unit** (**CPU**) is the electronic circuitry within a computer that executes instructions that make up a computer program. It is not designed to process data in parallel.

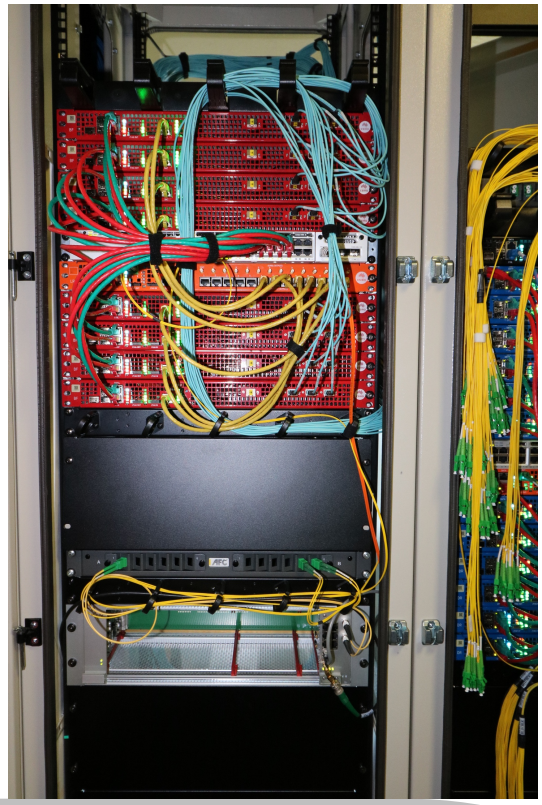
A **graphics processing unit** (**GPU**) has highly parallel structure so that they are more efficient than general-purpose CPUs for algorithms that process large blocks of data in parallel. It is a popular hardware to accelerate compute, but its power consumption is high and high-end GPUs are extremely expensive.

A **Field Programmable Gate Array (FPGA)** is an integrated circuit designed to be configured by a customer or a designer after manufacturing. Its power consumption is low, but it is hard to program.

Why in CASS do we use FPGAs?

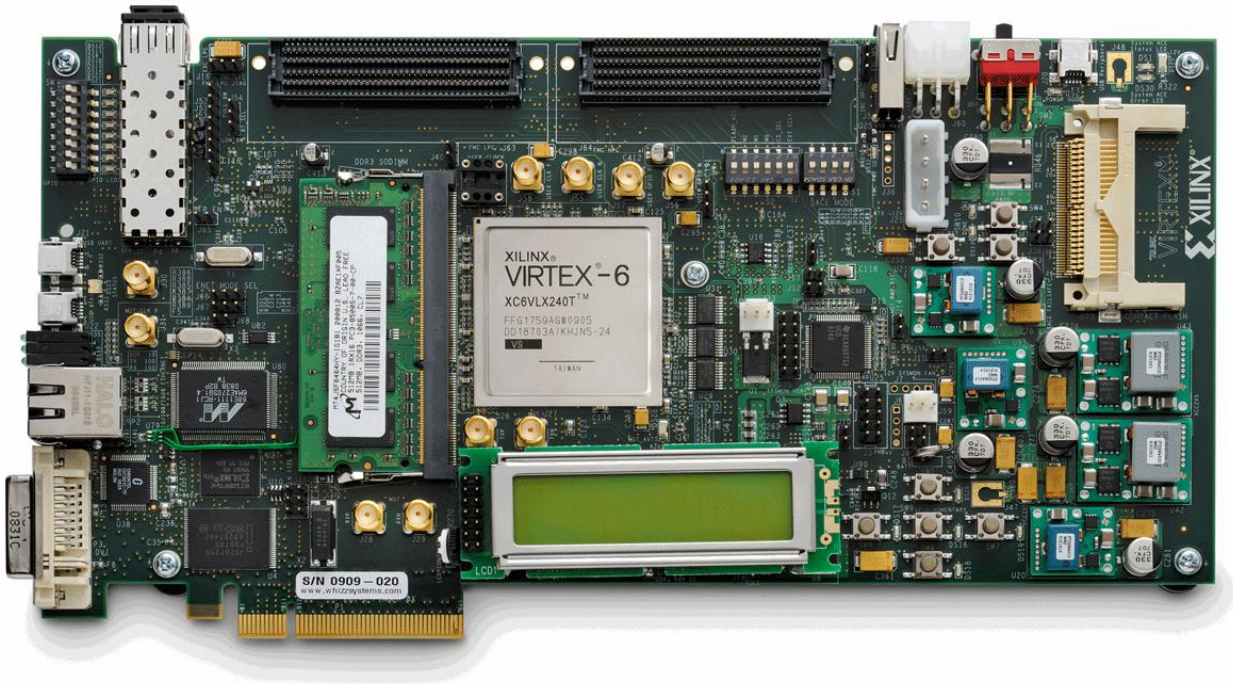
- They are mostly used in the telescope observing systems.
  - For example, digitizer, beamformer and correlator of ASKAP Phased Array Feed receiver





The problem with traditional FPGAs

- Need to be a hardware expert to program it.
- Need to arrange accessories connection
  - ➢ either buy it or build it yourself.



Xilinx

- **Xilinx**, Inc. (/ˈzaɪlɪŋks/ *ZY-links*) is an American technology company that develops highly flexible and adaptive processing platforms. The company invented the field-programmable gate array (FPGA).

What are Xilinx Accelerator Cards?

- Build on Xilinx FPGAs, for general developers;
- The first one was released at 16th Oct. 2018;
- It is not as popular as GPU yet;

- High performance (comparing with GPU):
  - More flexible on-chip memory;
  - Higher internal bandwidth;
  - Lower power consumption;

Popular Xilinx Accelerator Cards

	Feature	Alveo U200	Alveo U250	Alveo U280	Alveo U50
Dimensions	Width	Dual Slot	Dual Slot	Dual Slot	Single Slot
	Form Factor, Passive Form Factor, Active	Full Height, ¾ Length Full Height, Full Length	Full Height, ¾ Length Full Height, Full Length	Full Height, ¾ Length Full Height, Full Length	Half Height, ½ Length
Logic Resources [1]	Look-Up Tables	1,182K	1,728K	1,304K	872K
	Registers	2,364K	3,456K	2,607K	1,743K
	DSP Slices	6,840	12,288	9,024	5,952
DRAM Memory	DDR Format	4x 16GB 72b DIMM DDR4	4x 16GB 72b DIMM DDR4	2x 16GB 72b DIMM DDR4	–
	DDR Total Capacity	64GB	64GB	32GB	–
	DDR Max Data Rate	2400MT/s	2400MT/s	2400MT/s	–
	DDR Total Bandwidth	77GB/s	77GB/s	38GB/s	–
	HBM2 Total Capacity	–	–	8GB	8GB
	HBM2 Total Bandwidth	–	–	460GB/s	316GB/s[4]
Internal SRAM	Total Capacity	43MB	57MB	43MB	28MB
	Total Bandwidth	37TB/s	47TB/s	35TB/s	24TB/s
Interfaces	PCI Express®	Gen3 x16	Gen3 x16	Gen3 x16, 2xGen4 x8, CCIX	Gen3 x16, 2xGen4 x8, CCIX
	Network Interface	2x QSFP28	2x QSFP28	2x QSFP28	U50[2] - 1x QSFP28 U50DD[3] - 2x SFP-DD
Power and Thermal	Thermal Cooling	Passive, Active	Passive, Active	Passive, Active	Passive
	Typical Power	100W	110W	100W	50W
	Maximum Power	225W	225W	225W	75W
Time Stamp	Clock Precision	–	–	–	IEEE Std 1588
Tool Support	Vitis™ Developer Environment	Yes	Yes	Yes	Yes

Alveo™ Data Center Accelerator Cards

CSIRO

Comparing two Xilinx Accelerator cards with GPUs

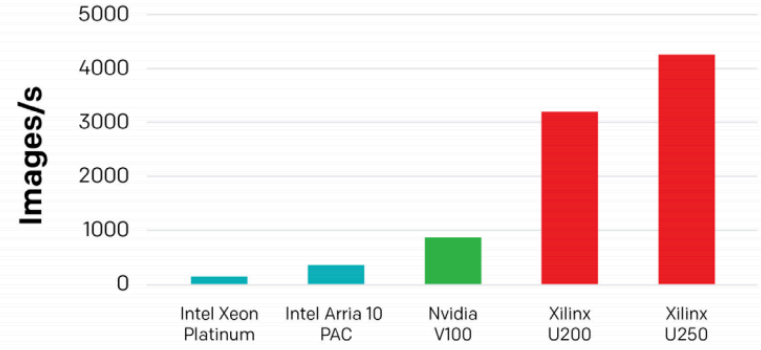
BENCHMARKS

Adapt and Accelerate Any Workload

AREA	PARTNER WORKLOAD	ALVEO ACCELERATION VS CPU
Database Search and Analytics	BlackLynx Unstructured Data Elasticsearch	90X
Financial Computing	Maxeler Value-at-Risk (VAR) Calculation	89X
Machine Learning	Xilinx Real-Time Machine Learning Inference	20X
Video Processing / Transcoding	NGCodec HEVC Video Encoding	12X
Genomics	Falcon Computing Genome Sequencing	10X

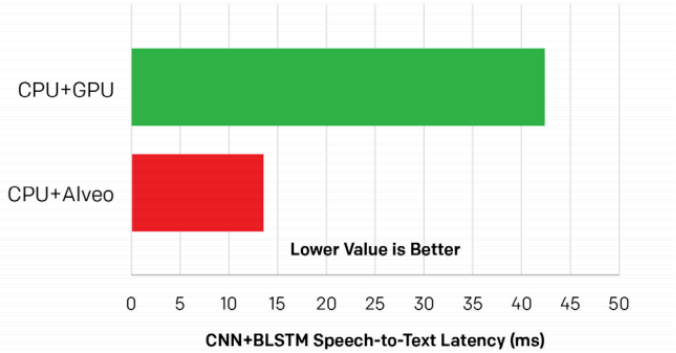
CPU Comparisons: Xeon c4.8xlarge AWS | Xeon E5-2643 v4 3.4GHz | Xeon Platinum c5.18xlarge AWS | Dual Socket E5-2680 v3 2.5GHz | Xeon f1.16xlarge

Increase Real-Time Machine Learning* Throughput by 20X



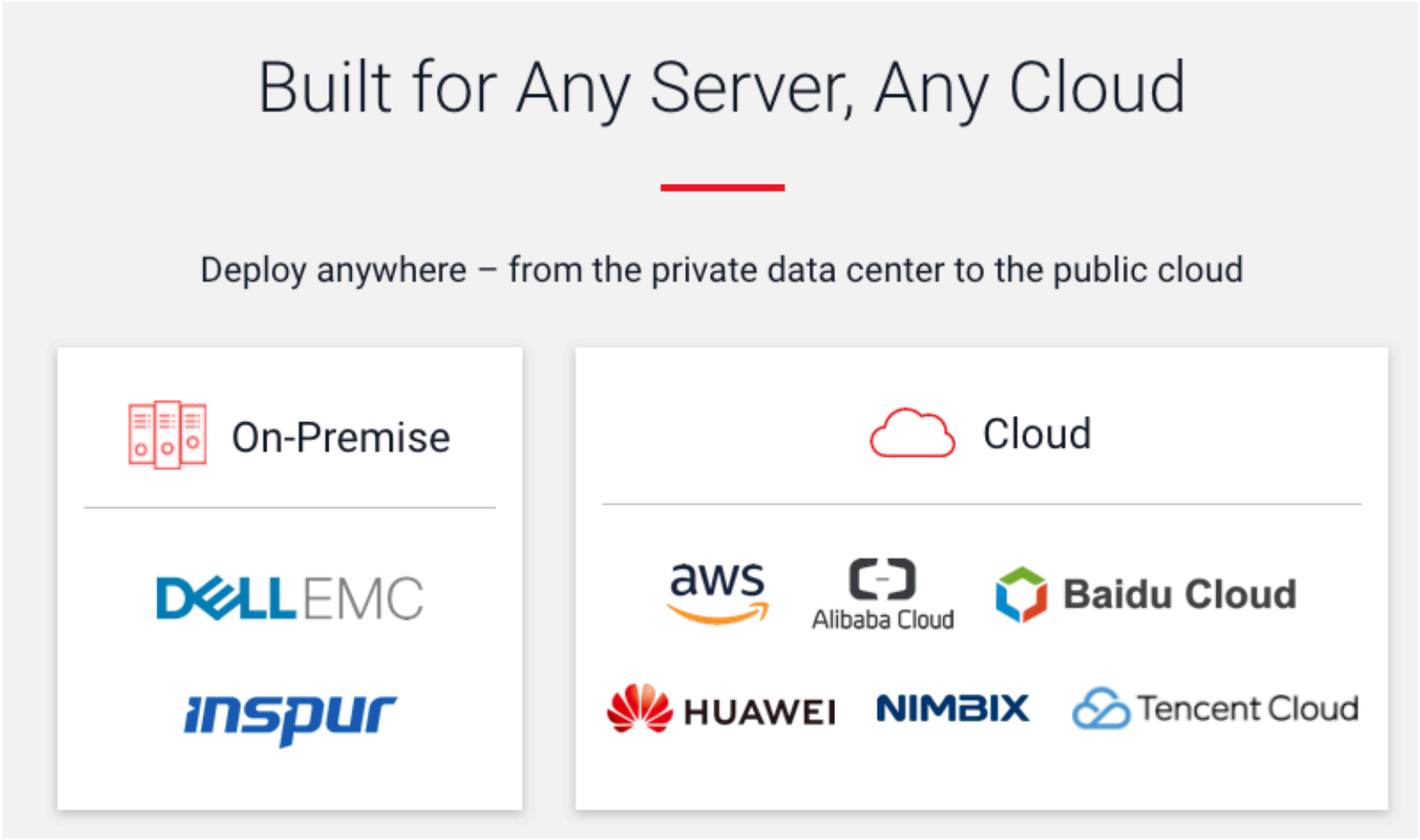
*GoogLeNet V1 : Accelerating DNNs with Xilinx Alveo Accelerator Cards White Paper

Reduce ML Inference Latency by 3X



CPU+GPU: Nvidia P4 + Xeon CPU E5-2690 v4 @2.60GHz (56 Cores)
CPU+Alveo: Alveo U200 or U250 + Xeon CPU E5-2686 v4 @2.3GHz (8 Cores)

Who is using these new cards?

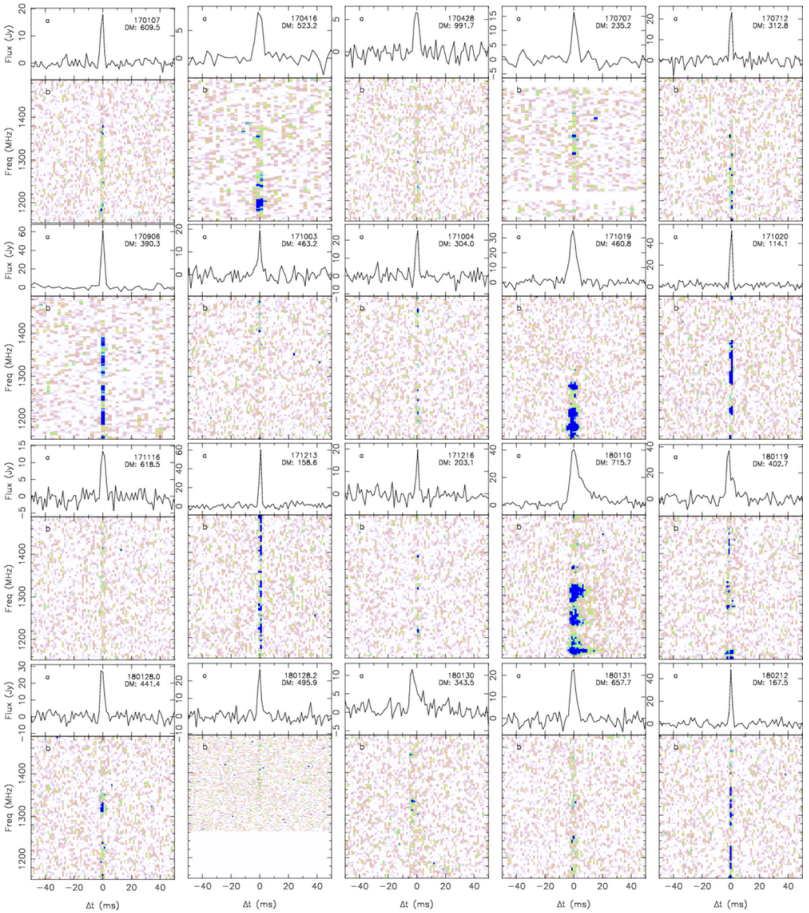


Why is CASS interested in the new card?

- It is easier to program:
  - No hardware description language required
- It is ready to use:
  - Plug it into a computer and use it.

Why I am interested in this card?

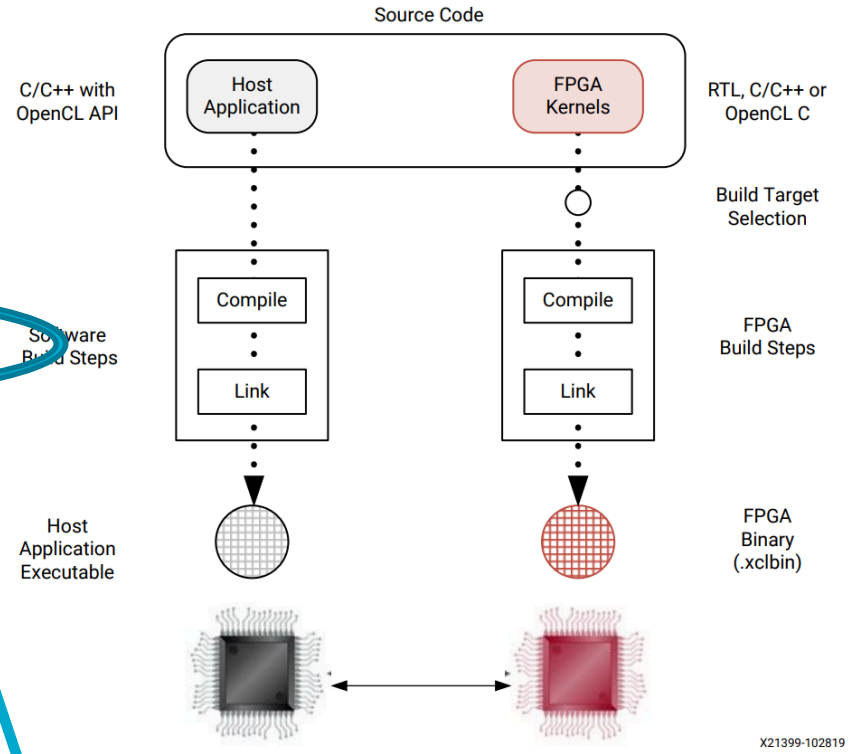
- GPU's on-chip memory is not flexible to run ASKAP coherent fast-radio-burst detection pipeline;
- Get experience on the new card to see if we can develop telescope signal processing system with it in future.



How to develop with it?

Development methodology

- Like GPU approach:
  - Host does memory management, communication control, etc.
  - Kernel does accelerations;
  - PCIe for data transfer;
- Host in C/C++/OpenCL;
- Kernel in RTL, C/C++ or OpenCL C;
  - RTL is unnecessary for kernel programming;
  - It is an option kept for traditional FPGA developers;
- General libraries provided;



Programming languages

Development iteration

- **Software emulation:** kernel runs on CPU, takes couple of minutes to compile and execute;
- **Hardware emulation:** kernel runs on CPU, but emulation the execution on hardware, takes about ten minutes to compile, couple of hours to execute;
- **Hardware execution:** kernel runs on FPGA, takes couple of hours to compile, take seconds to execute;

Software Emulation	Hardware Emulation	Hardware Execution
Host application runs with a C/C++ or OpenCL model of the kernels.	Host application runs with a simulated RTL model of the kernels.	Host application runs with actual hardware implementation of the kernels.
Used to confirm functional correctness of the system.	Test the host / kernel integration, get performance estimates.	Confirm that the system runs correctly and with desired performance.
Fastest build time supports quick design iterations.	Best debug capabilities, moderate compilation time with increased visibility of the kernels.	Final FPGA implementation, long build time with accurate (actual) performance results.

Acceleration with FPGA is easy

#PRAGMA HLS XXX

is all you need to know …

Instruction level parallelism

A slide for experts!

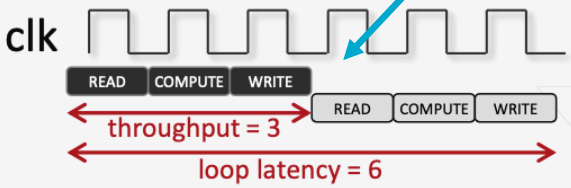
One iteration at a time without pragma

```
void F (int A[2],int Z[2]) {
# PRAGMA HLS PIPELINE

add: for (i=1;i=<2;i++) {
       Z[i] = A[i] + 10;
     }
}
```

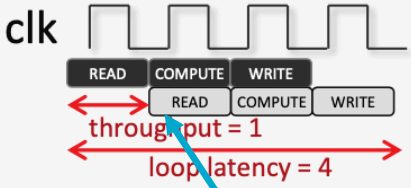
Func Pipelining

Standard C code

default



PIPELINE



Start the next iteration before finishing the previous one!

PRAGMA provides additional information to the compiler

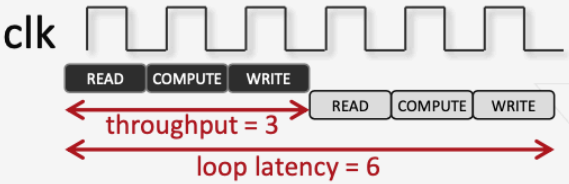
Instruction level parallelism

A slide for experts!

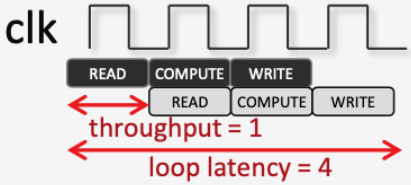
```
void F (int A[2],int Z[2]) {
# PRAGMA HLS PIPELINE

add: for (i=1;i=<2;i++) {
       Z[i] = A[i] + 10;
     }
}
```

Func Pipelining

default

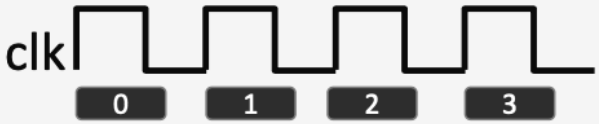


PIPELINE

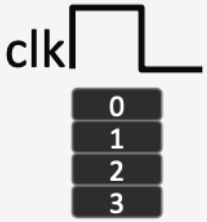



```
void F (...)  {
...
add: for (i=0;i<=3;i++) {
#pragma HLS UNROLL
  b = a[i] + c;
...
```

Default: 4 cycles



Unroll: 1 cycle

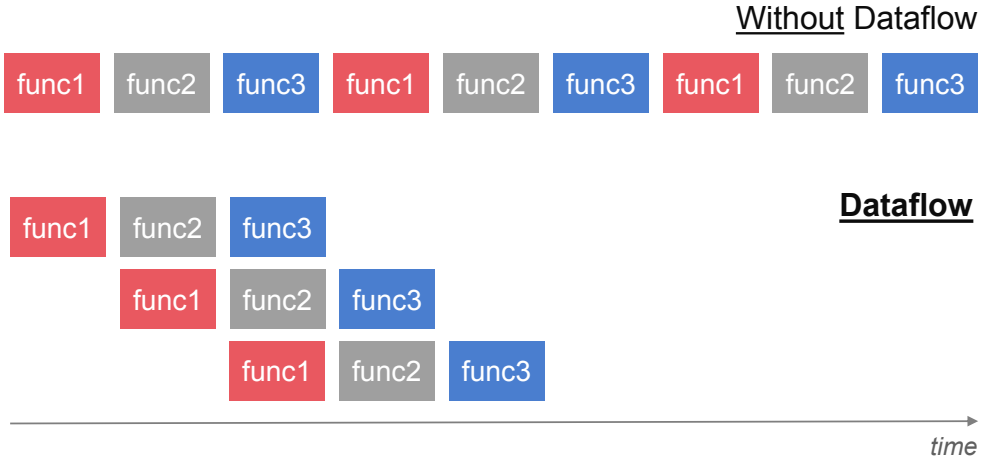


Slide taken from Xilinx

A slide for experts!

Task-Level Parallelism

```
for (int i=0; i<N; i++)
{
  #pragma HLS DATAFLOW
  func1();
  func2();
  func3();
}
```



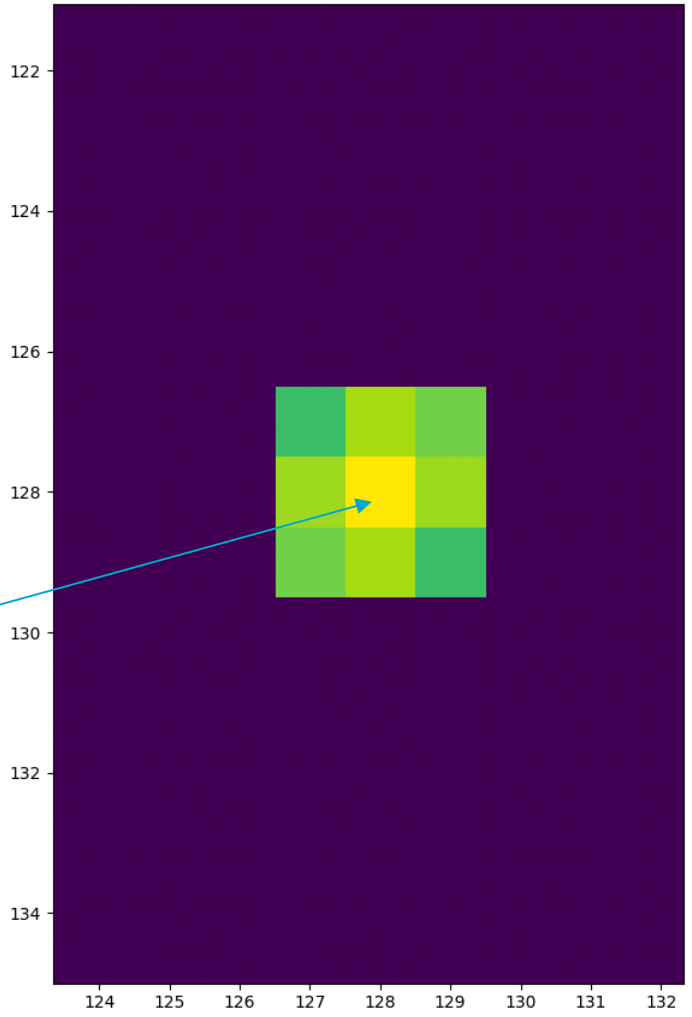
- Create custom dataflow pipelines
- Multiple tasks executing simultaneously
- Streaming programming paradigm

Sometimes acceleration may not be that easy

Not going to detail here…

ASKAP coherent FRB detection pipeline

- Developing a pipeline with Xilinx accelerator card U280
- Expect that this will improve ASKAP FRB detection sensitivity by a factor of 10



Simulated FRB candidate coming through the pipeline

FPGAs vs GPUs for astronomy data processing

- Should future systems be FPGAs or GPUs based?
- With experience in both GPU and FPGA programming and my thoughts are:
  - GPUs are relatively easy to program and faster to iterate your code;
  - New Xilinx system much easier than previous FPGAs, but still in development;
  - It is worth to follow and use these developments as:
    - ✓ It is a more cost-efficient solution (both power consumption and price is lower than GPUs);
    - ✓ Its flexible on-chip memory and integrated network interface are very attractive;

Want to try it out?

- Xilinx Nimbix Cloud has Xilinx Vitis and accelerator card installed https://www.xilinx.com/xilinxtraining/assessments/portal/alveo/intro_nimbix_cloud/story.html, free trial is available.

- Come and chat with me. xinping.deng@csiro.au, room 84.

Thank you

Astronomy and Space Science
Xinping Deng
Research Engineer
E xinping.deng@csiro.au