Software-like
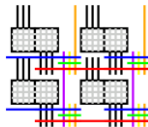# Incremental Refinement on FPGA ?

Dongjoon Park, 박동준
dopark@seas.upenn.edu

Implementation of Computation Group
University of Pennsylvania
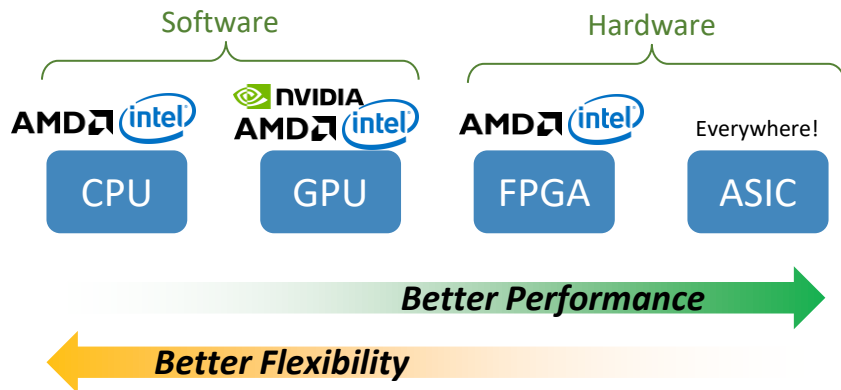
Penn
Engineering
UNIVERSITY of PENNSYLVANIA

Software-like Incremental Refinement on FPGA

How many of you have heard of "**FPGAs**"?
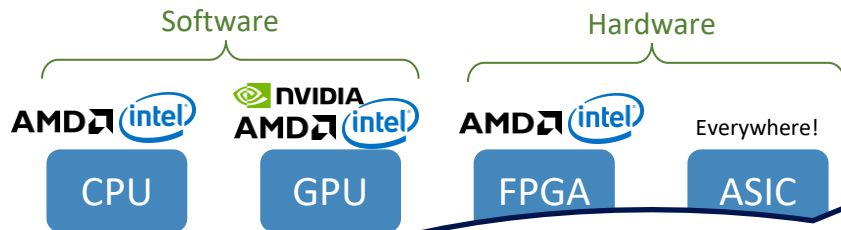
Software-like Incremental Refinement on FPGA

- FPGA: Field-Programmable Gate Arrays
- Compare with other hardware platforms…

+ Good at computationally intensive task
+ Flexible; can be reprogramed after manufacturing

Software-like Incremental Refinement on FPGA

- FPGA: Field-Programmable Gate Arrays
- Compare with other hardware platforms…

Software   Hardware

AMD intel   NVIDIA AMD intel   AMD intel   Everywhere!

| CPU | GPU | FPGA | ASIC |

My friend

If FPGA's that good, then why not using FPGAs everywhere?

+ Good at computationa~
+ Flexible; can be reprogramed after ~~manufac~~turing

Software-like Incremental Refinement on FPGA

- Problem
  - FPGA <u>compilation</u> takes forever

The process of translating the user code (C, C++, Verilog) in a form that hardware can interpret and execute

User code

```
assign a = b + c;
…
```

Some magic…

Program your chip

- CPU, GPU (Software): milliseconds, seconds, minutes
- FPGA (Hardware): **minutes, hours, days**
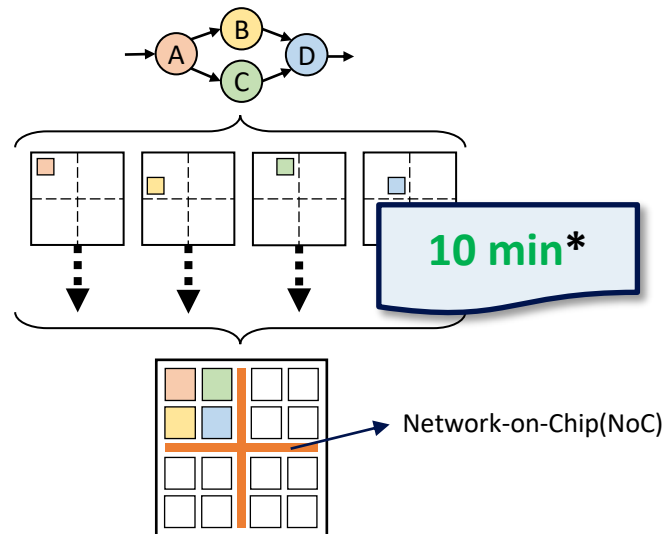
Software-like Incremental Refinement on FPGA

- Problem
  - FPGA <u>compilation</u> takes forever

  - You have a design… Wait for an hour to test it on FPGA
  - … and you discover a small change you need to make
  - Now, you must wait *another* hour to test the modified design? ☹

Software-like Incremental Refinement on FPGA

- Idea: Fast separate compilations on FPGA
  - Divide-and-conquer strategy!

1 hour*

10 min*

Network-on-Chip(NoC)
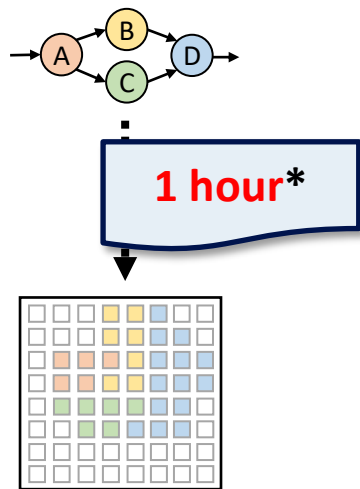
Vendor tool(from AMD, Intel)'s slow monolithic compilation

Fast separate compilations in parallel

*Example values

Software-like Incremental Refinement on FPGA

- Idea: Fast separate compilations on FPGA
  - Divide-and-conquer strategy!
  - <u>Incremental Refinement</u> in SW
    - We never come up with the final design at one shot
    - Start from something that's barely functional… then incrementally add functionality
    - Not possible in FPGA because each design iteration takes so long
  - With our strategy, SW-like Incremental Refinement on FPGA is possible! ➔ Better design productivity!

Monolithic
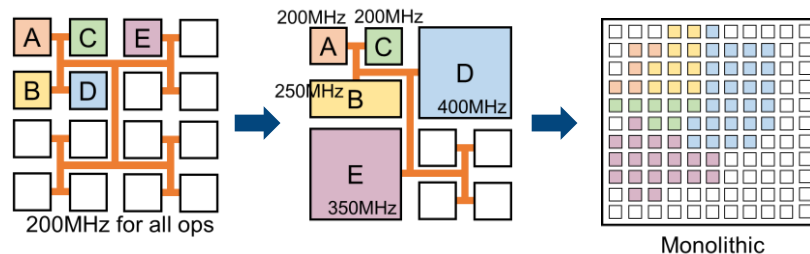
Software-like Incremental Refinement on FPGA

- **Broader Impact? (in high-level)**
  – e.g. AI Chips

- Hardware development is time-consuming…
- With our modular design methodology, it can be accelerated!
➔ Better cell phone, laptop, ChatGPT, everything!

Tensor Processing Unit products[13][14][15]

	TPUv1	TPUv2	TPUv3	TPUv4[14][16]	TPUv5[17]	Edge v1
Date introduced	2016	2017	2018	2021	2023	2018
Process node	28 nm	16 nm	16 nm	7 nm	Unstated	
Die size (mm^2)	331	< 625	< 700	< 400	Unstated	
On-chip memory (MiB)	28	32	32	32	48	
Clock speed (MHz)	700	700	940	1050	Unstated	
Memory	8 GiB DDR3	16 GiB HBM	32 GiB HBM	32 GiB HBM	16 GB HBM	
Memory bandwidth	34 GB/s	600 GB/s	900 GB/s	1200 GB/s	819 GB/s	
TDP (W)	75	280	220	170	Not Listed	2
TOPS (Tera Operations Per Second)	23	45	123	275	393	4
TOPS/W	0.31	0.16	0.56	1.62	Not Listed	2

Kind of slow…

<Google TPU products[1]>

Thank you ☺

[1] Wikipedia, "Tensor Processing Unit", https://en.wikipedia.org/wiki/Tensor_Processing_Unit (accessed Apr. 8th, 2024)

Penn Engineering
UNIVERSITY of PENNSYLVANIA