



Workshop: Accelerating AI – Challenges and Opportunities in Cloud and Edge Computing

February 18, 2020, Toronto, ON, Canada



Towards an Accelerated and Power-Efficient Mobile DNA Sequencing Device: An Edge-Computing Perspective

Presenter: Dr. Karim Hammad

Supervisor: Prof. Sebastian Magierowski

**Electrical Engineering & Computer Science
York University
Toronto, Canada**



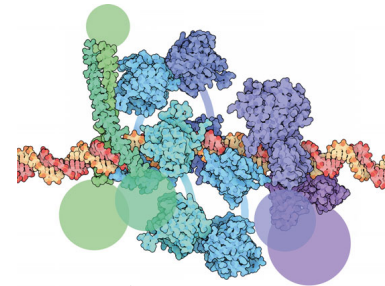
Agenda

- Background
- 3rd Generation Nanopore DNA Sequencing
- Mobile DNA Sequencing Trends
- FPGA Acceleration of HMM Basecalling
- Neural Network Basecalling
- Future Embedded Basecalling Solution
- Summary

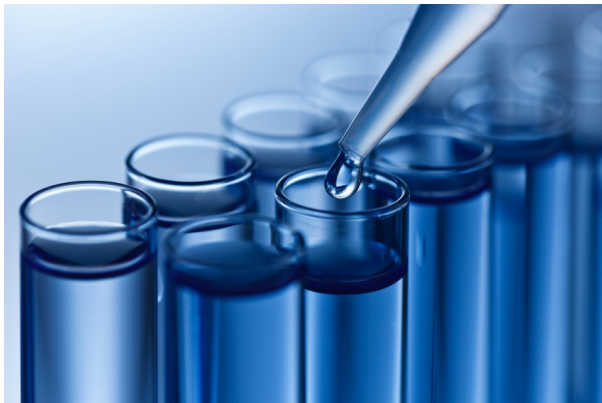
Background: DNA Sequencing



- **Bioinformatics**
 - computers + biological data (NIH)
- more narrowly...
 - analysis of biomolecules
 - their make-up, structure, and function
 - proteins, DNA, RNA, etc.



[Goodsell'16,©Springer]



- **“Applications”**
 - Food safety
 - Plant health
 - Animal health
 - Cancer research

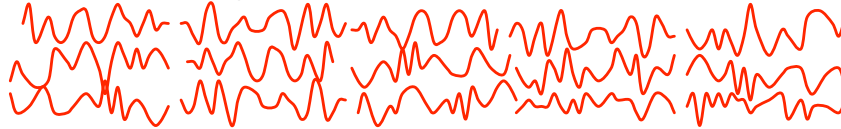
Background: DNA Sequencing Pipeline

1. DNA Isolation

2. DNA fragmentation

3. DNA (chemical) amplification

4. DNA-to-signal transduction



5. Basecalling

ACCTGTCGT GCAAAAATC TCAAAACGG CAAATGCGC ACGGACGGT
GGCACATAA AGTGCAACC CCAATTTAC CTAGATTAC CCTTGAGAA
TCTAGTCAT CCTAATGC TCTCCCGAG CTGTGTCAT TTCCGCAC

6. Alignment



7. Sorting, 8. De-duplication,
9. Local align, 10. Quality score adjustment
11. Variant calling

wet lab

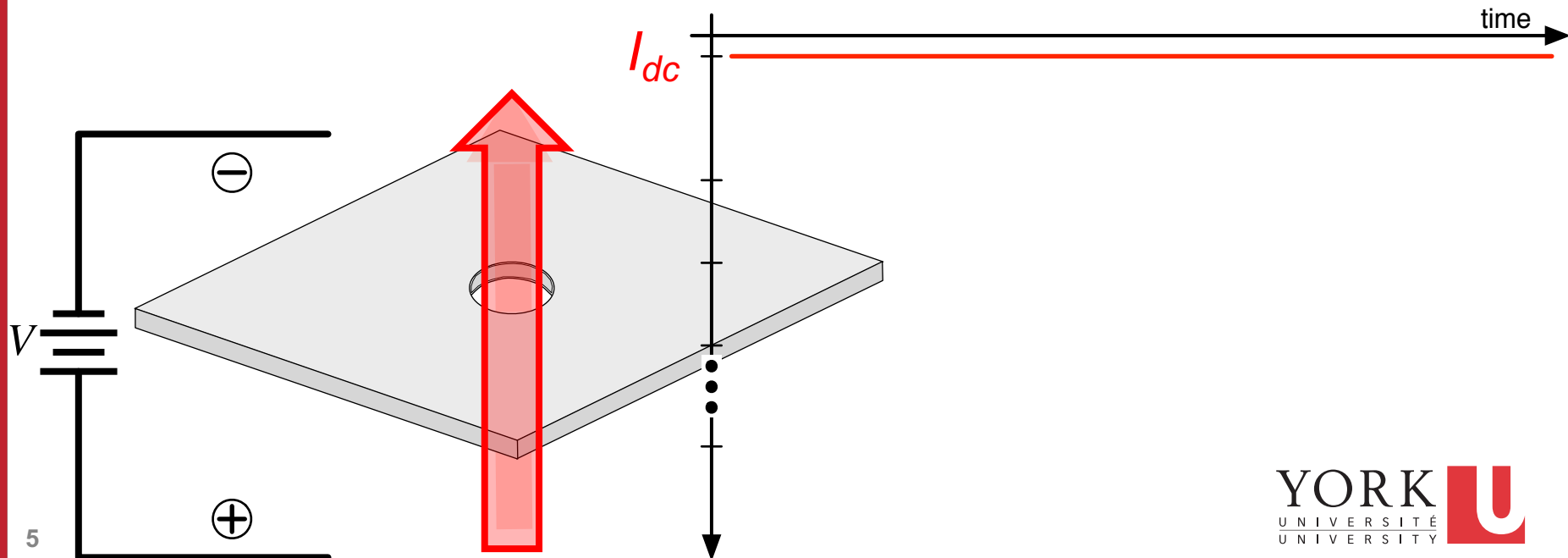
sequencing machines

computers



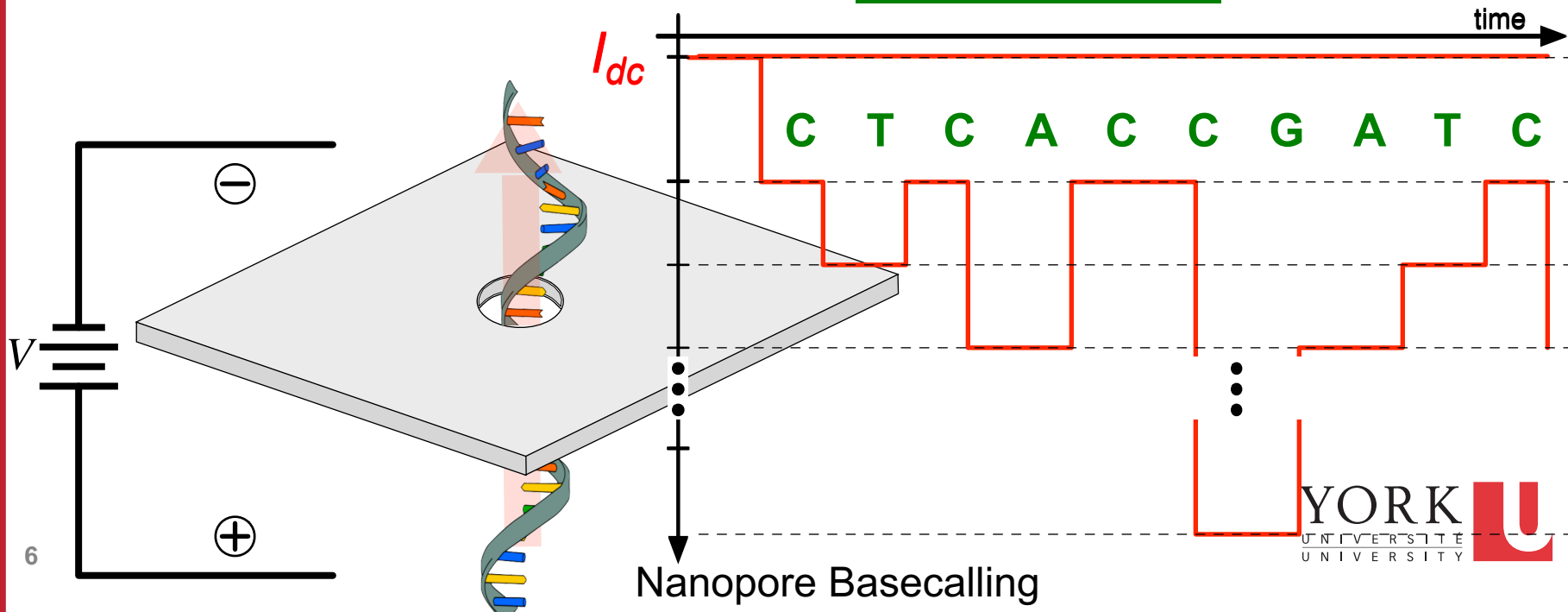
3rd Gen. Nanopore DNA Sequencing

- Miniaturization led by nanopore technology
- A little review:
 - nanopore: small hole ~ 2 nanometers
 - immerse in conductive fluid
 - apply DC voltage → get DC current

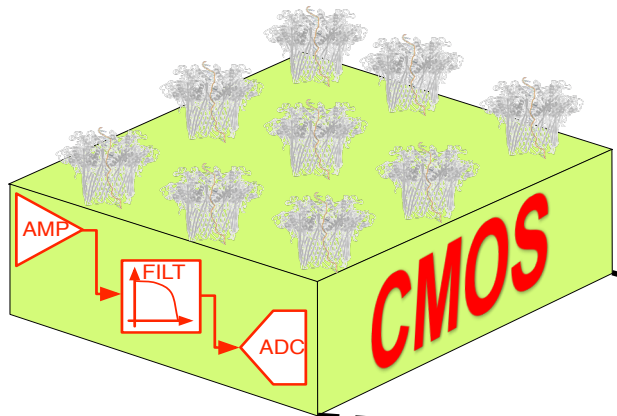


3rd Gen. Nanopore DNA Sequencing

- Miniaturization led by nanopore technology
- A little review:
 - As DNA goes through, DC current is modulated
 - Modulations proportional to DNA bases going through pore
 - Convert modulation to bases: **BASECALLING**



Mobile DNA Sequencing Market Trends



Under development



[© Oxford Nanopore Tech]

2015



[© Oxford Nanopore Tech]

Ubiquitous sequencing vision

Since 2005



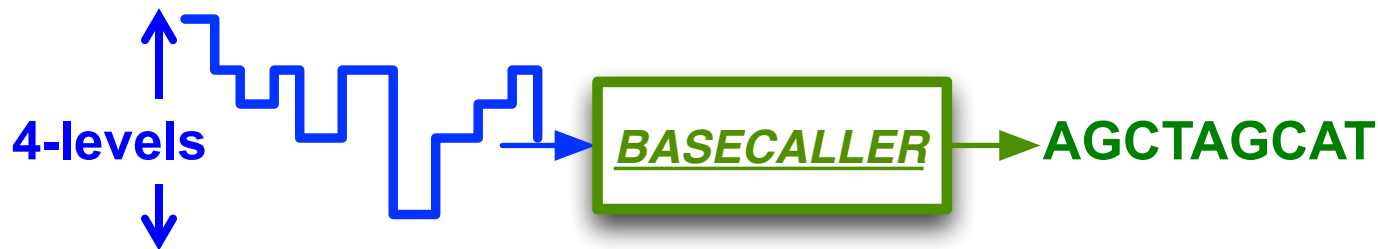
220 kg
1,500 W
\$1,000,000
~10 genomes/hr

0.1 kg
5 W
\$1,000
~0.13 genomes/hr

[© Illumina, Inc]

Acceleration of HMM Basecalling: Background

- **Ideally:** 4 levels mapped to **A, C, G**, or **T**

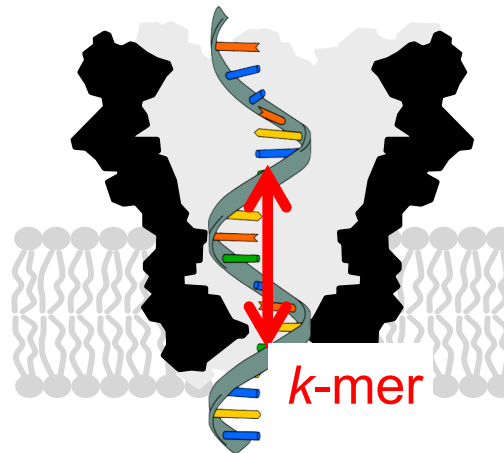


- **Really:** More than one base in pore at a time

- $k \approx 3-6$ bases

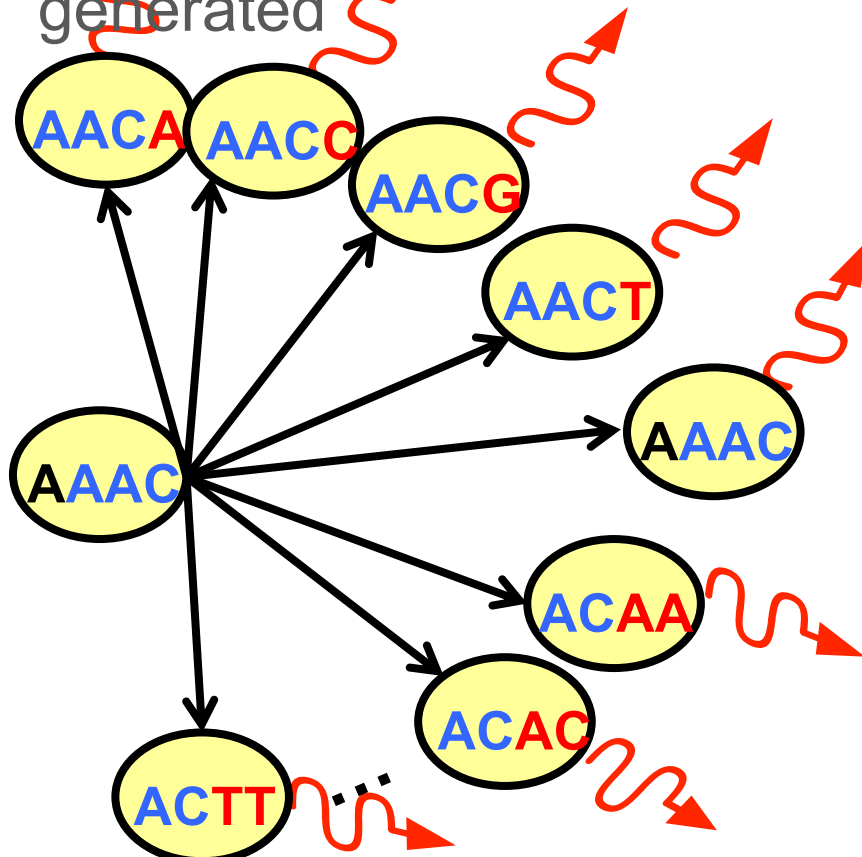
- k -mer

- 3-mer
- 4-mer
- 5-mer
- 6-mer



Acceleration of HMM Basecalling: Background

- Basecallers need...
- ...model of how signal is generated

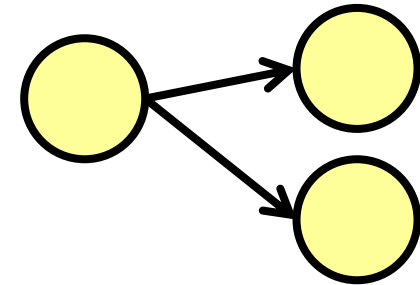


- Transition model

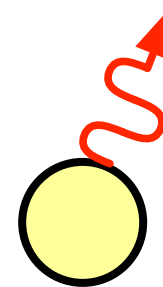
- step

- stay

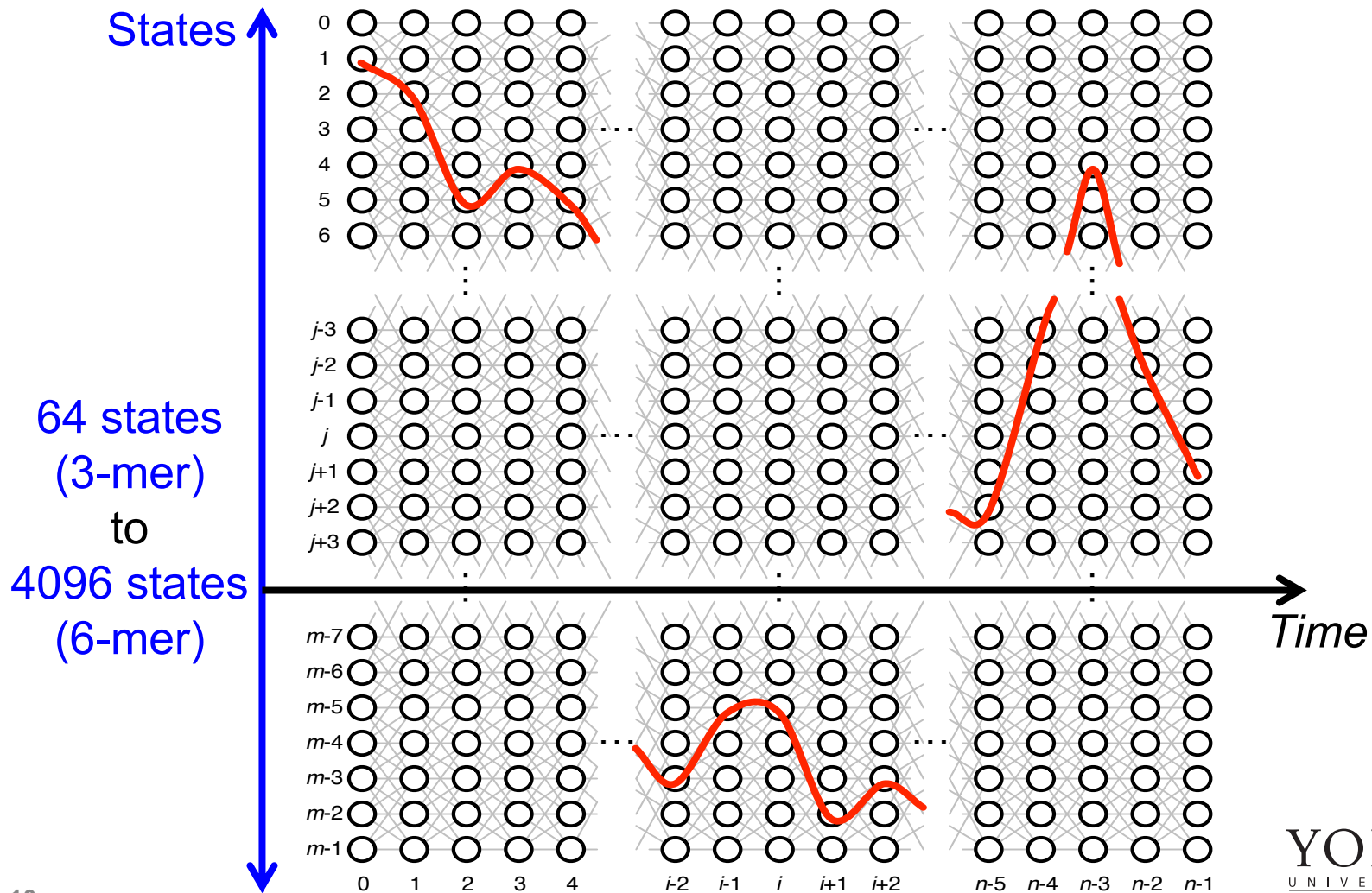
- skip



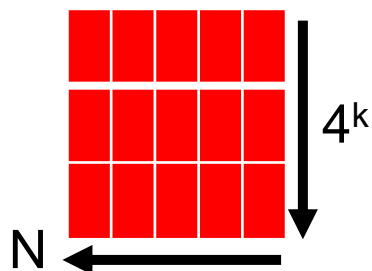
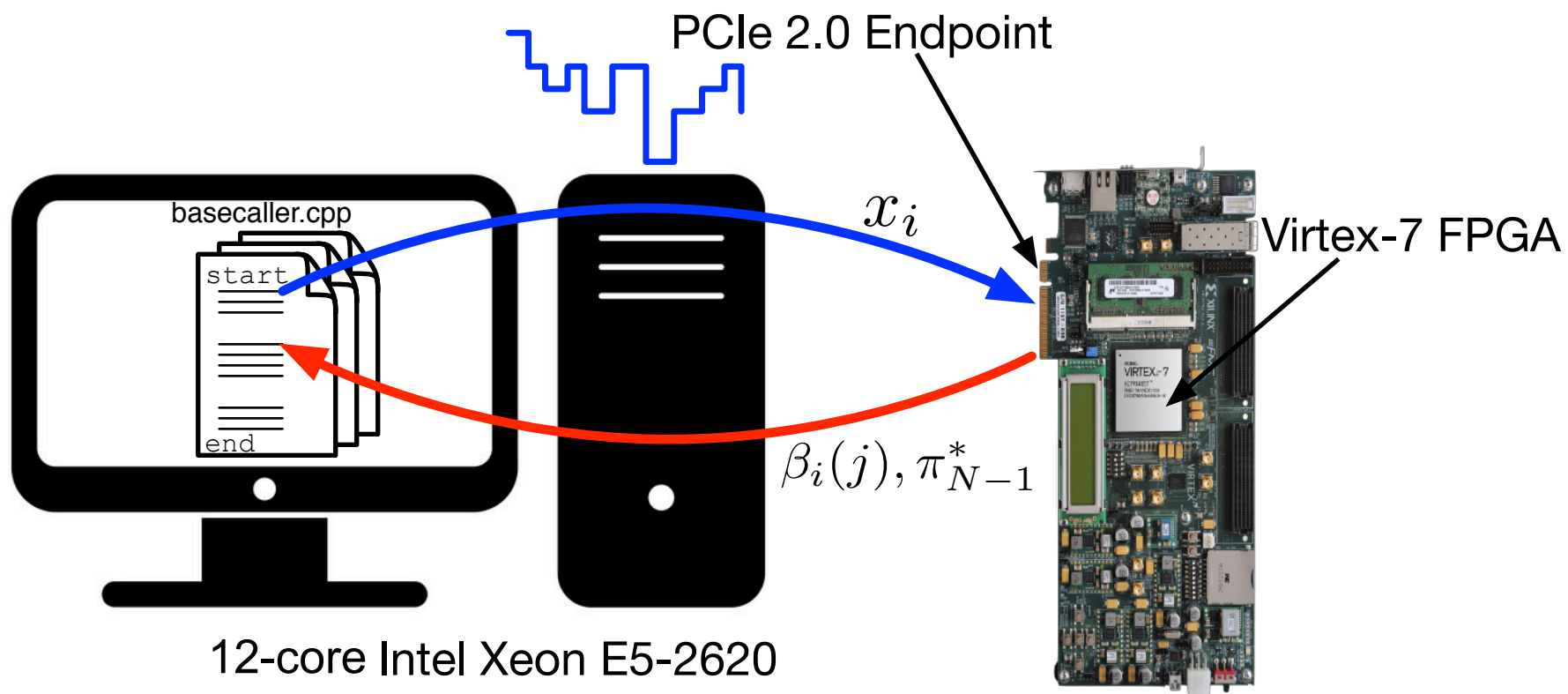
- Emission Model



Acceleration of HMM Basecalling: Background



Acceleration of HMM Basecalling: Test Setup



Acceleration of HMM Basecalling: 3-mer Architecture

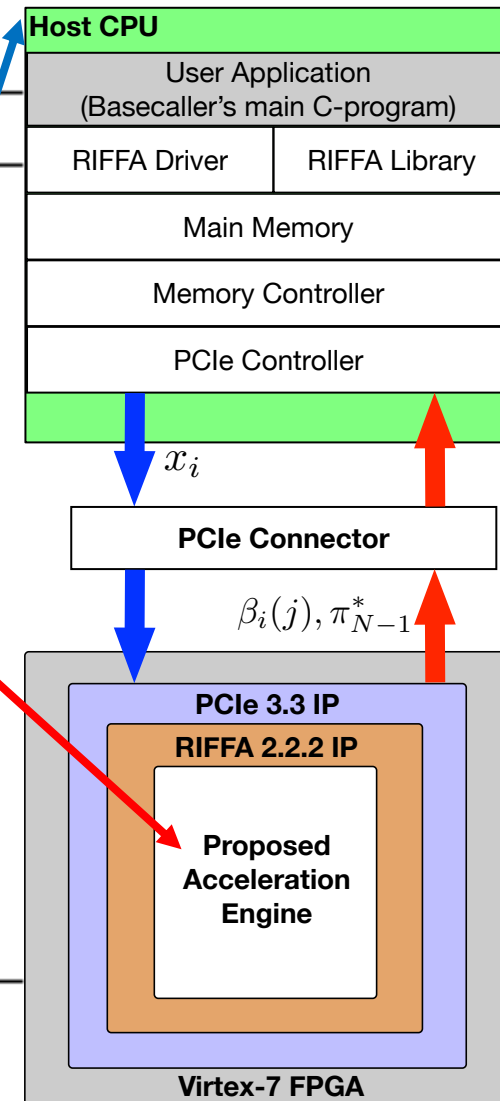
Algorithm 1 HMM Basecalling Algorithm

```

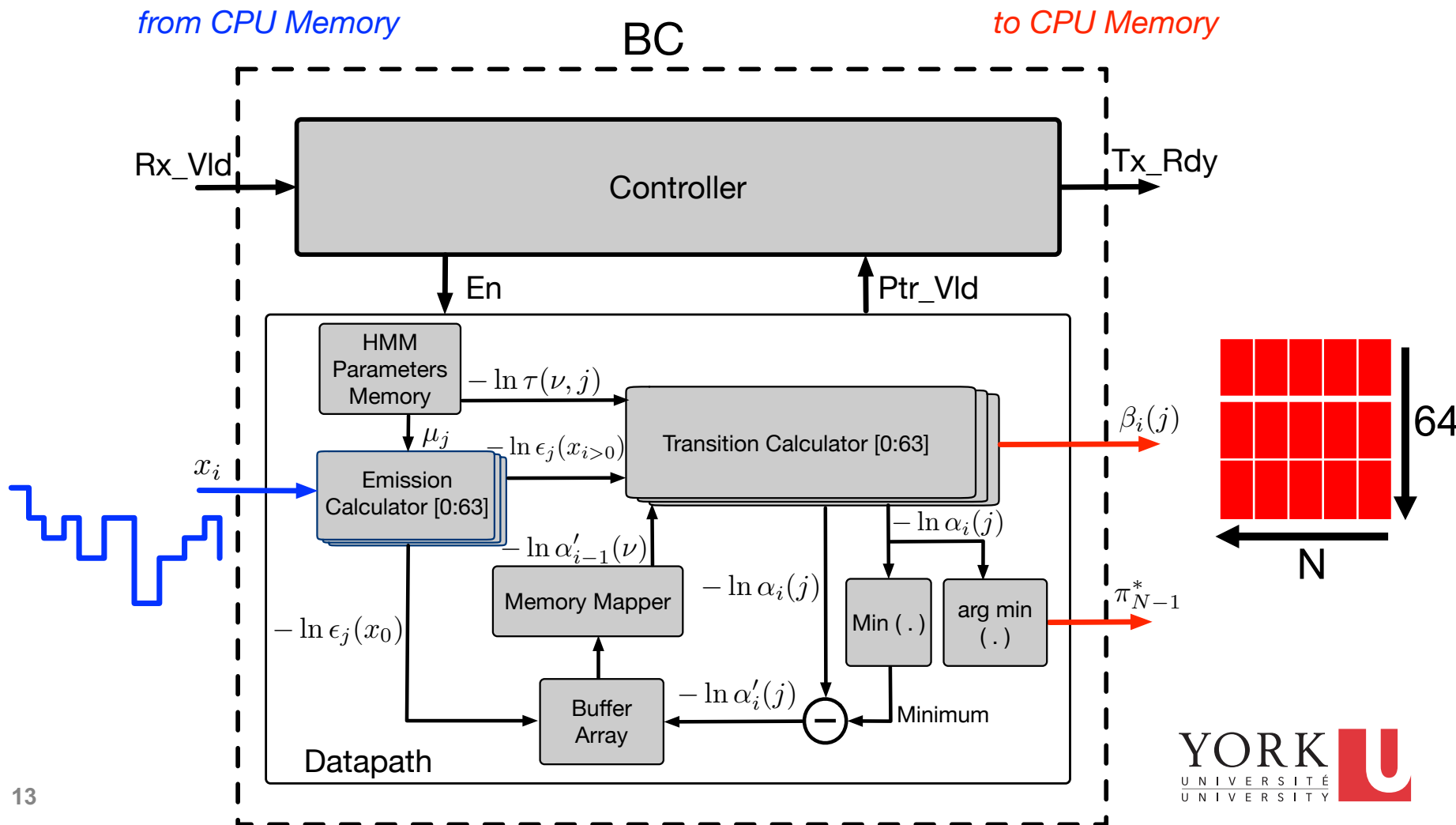
1: Initialize (on FPGA):
2:  $\alpha_0(j) \leftarrow \epsilon_j(x_0) \forall j \in \{0, \dots, 63\}$ 
3: Iterate (on FPGA):
4: for  $i \leftarrow 1, N - 1$  &  $\forall j$  do
5:    $\alpha_i(j) \leftarrow \epsilon_j(x_i) \max_{\nu \in \omega(j)} [\alpha_{i-1}(\nu) \tau(\nu, j)]$ 
6:    $\beta_i(j) \leftarrow \arg \max_{\nu \in \omega(j)} [\alpha_{i-1}(\nu) \tau(\nu, j)]$ 
7: end for
8: EndState (on FPGA):
9:  $\pi_{N-1}^* \leftarrow \arg \max_j [\alpha_{N-1}(j)]$ 
10:  $\hat{a}_{N-1} \leftarrow \pi_{N-1}^* \& 3$ 
11: Traceback (on CPU):
12: for  $i \leftarrow N - 1$  to 1 do
13:    $\pi_{i-1}^* \leftarrow \beta_i(\pi_i^*)$ 
14:    $\hat{a}_{i-1} \leftarrow \pi_{i-1}^* \& 3$ 
15: end for
  
```

64-states Trellis

Traceback



Acceleration of HMM Basecalling: 3-mer Architecture



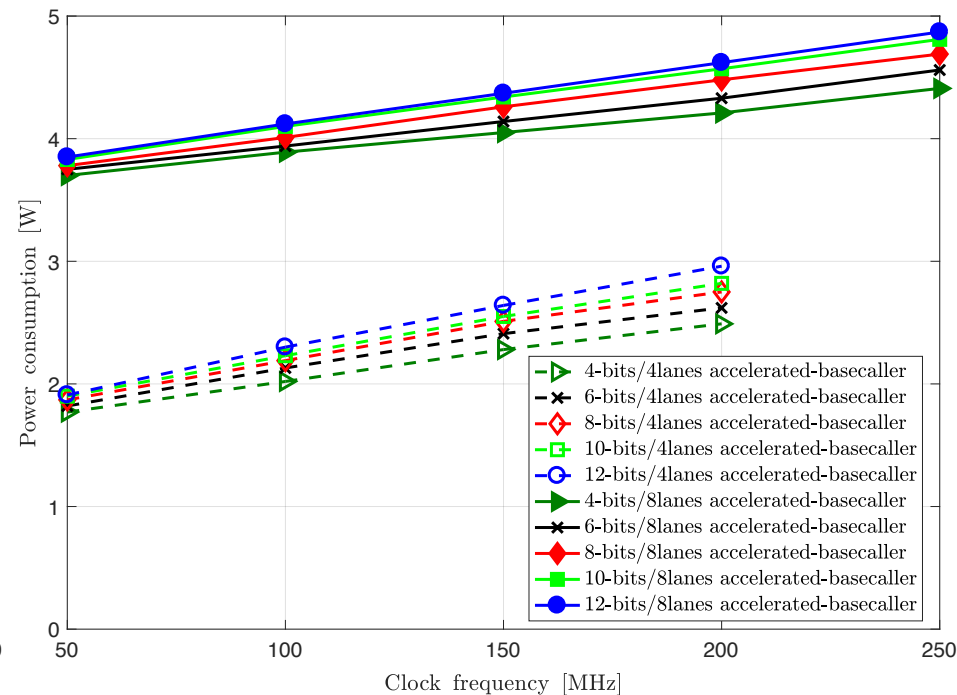
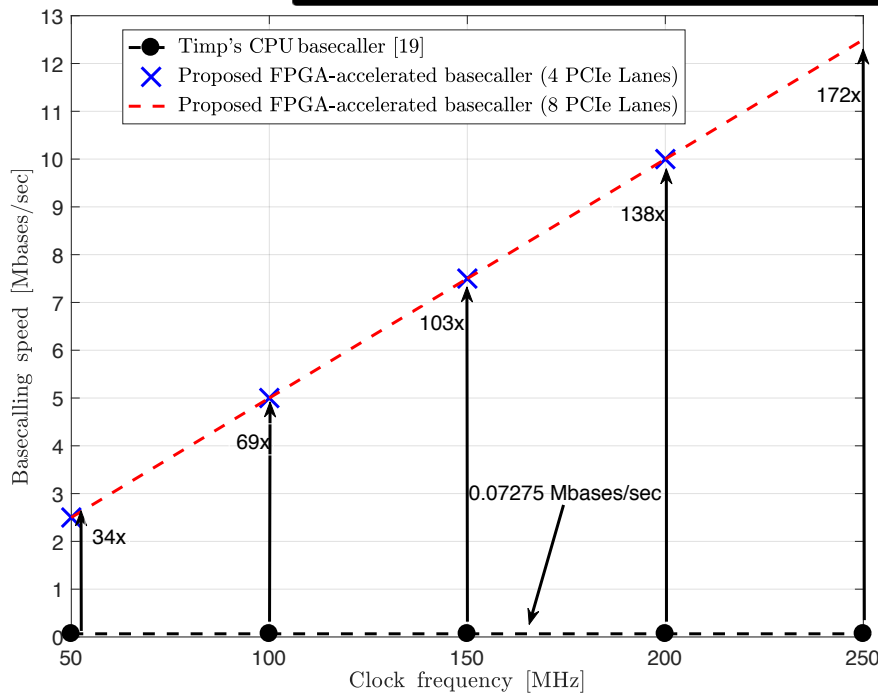
Acceleration of HMM Basecalling: 3-mer Results

IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS, VOL. 14, NO. 1, FEBRUARY 2020

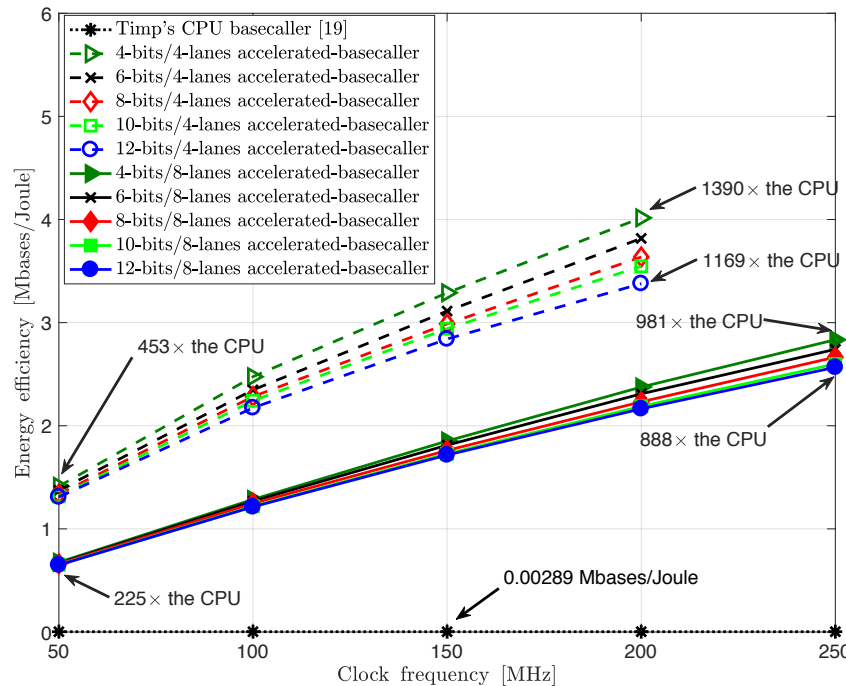
65

FPGA-Accelerated 3rd Generation DNA Sequencing

Zhongpan Wu, *Member, IEEE*, Karim Hammad , *Member, IEEE*, Ebrahim Ghafar-Zadeh , *Member, IEEE*, and Sebastian Magierowski, *Member, IEEE*

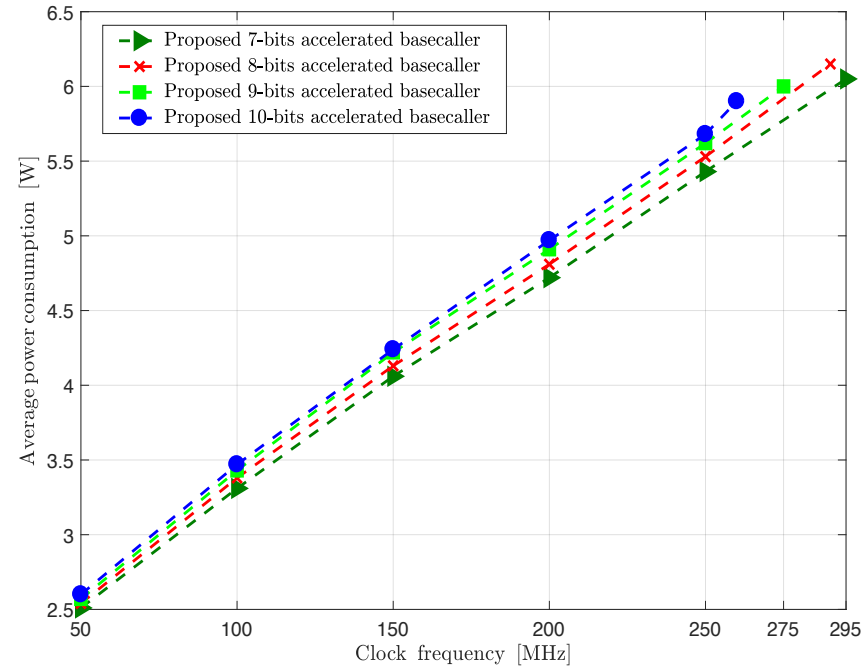
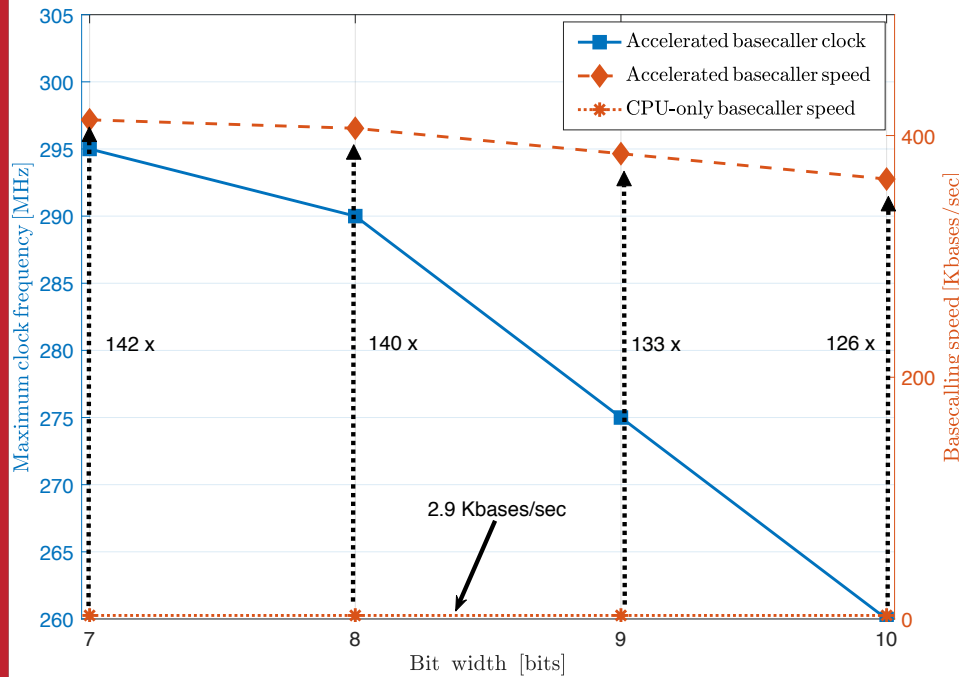


Acceleration of HMM Basecalling: 3-mer Results



	4-PCIe Lanes				8-PCIe Lanes			
	LUT	FF	DSP	ED	LUT	FF	DSP	ED
4-bits	12%	6%	53%	1.47	12%	6%	53%	1.76
6-bits	14%	7%	55%	1.66	15%	7%	55%	1.95
8-bits	17%	8%	55%	1.72	18%	8%	55%	2.0
10-bits	20%	9%	57%	1.87	21%	9%	57%	2.05
12-bits	23%	10%	59%	1.91	24%	11%	59%	2.15

Acceleration of HMM Basecalling: 6-mer Results



Acceleration of HMM Basecalling: 6-mer Results

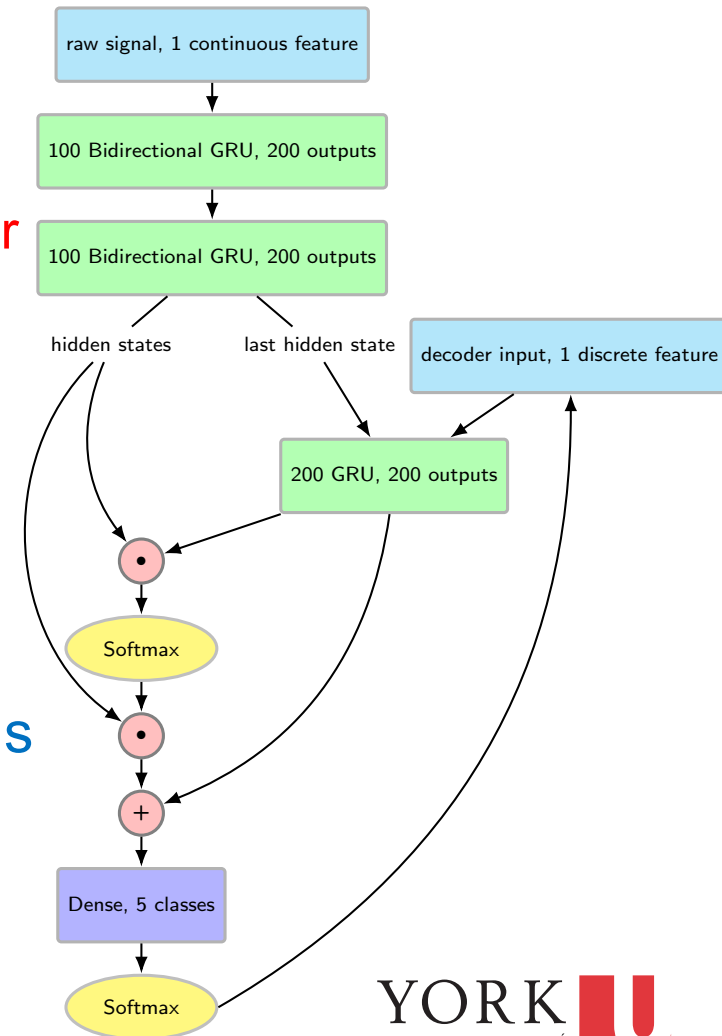
Bit width	LUT	FF	DSP	BRAM
7-bits	19%	9%	2.3%	74%
8-bits	21%	10%	2.3%	74%
9-bits	23%	11%	2.3%	74%
10-bits	25%	12%	4.6%	74%

Neural Network Basecalling

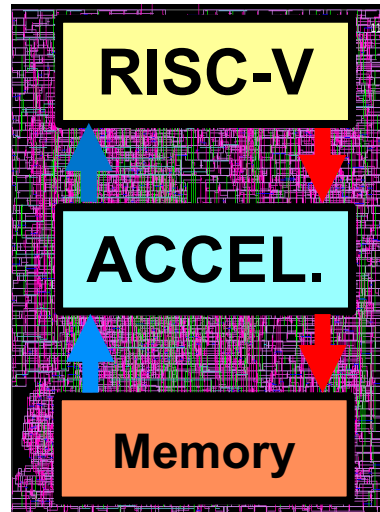
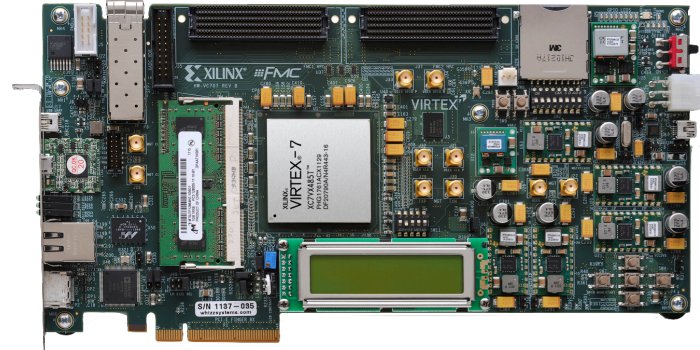
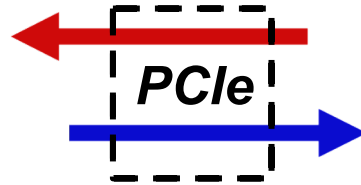
- Hidden Markov Model (HMM)
 - ~ 55 GFLOP/s
 - ~70% accuracy
 - Note: sequencers can run 1000X faster
 - $55 \times 1000 = 55$ TFLOP/s
- Neural Network (NN)
 - ~ 200 GFLOP/s
 - ~90% accuracy
 - Note: sequencers can run 1000X faster
 - $200 \times 1000 = 200$ TFLOP/s

Neural Network Basecalling

- Our developed NN basecaller
 - ❑ encoder-decoder structure
 - convert time series to context vector
 - convert context vector to labels
 - ❑ employ attention mechanism
 - context vector filter
 - ❑ ~350,000 parameters
 - ❑ 90% accuracy
 - ❑ competitive with much larger designs
 - e.g. 2,000,000 parameter Chiron
 - 3e-3 genomes/hour on GTX 1080
 - 40X too slow for real-time



Future Embedded Basecalling Solution



Summary

- Mobile DNA sequencing is an evolving and challenging market.
- FPGA acceleration demonstrated an attractive performance for an HMM-based sequencer (~142X faster than a 12 core Intel Xeon CPU @ 6W power margin).
- Fully embedded ASIC solution would offer a great potential for a market competition.

Thank You

Questions?