New Frontiers in On-Chip Image Processing

Richard Younger

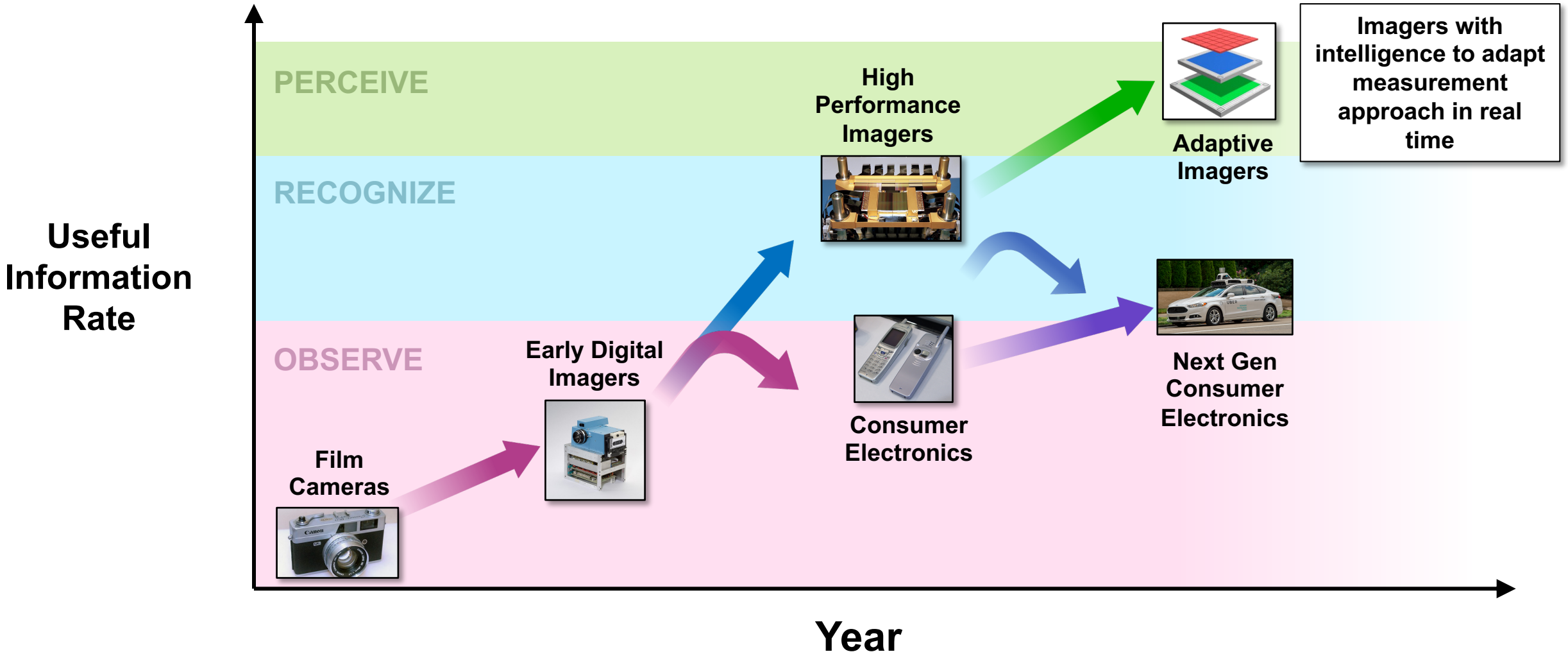
CPAD Instrumentation Frontier Workshop 2021

March 18-21, 2021

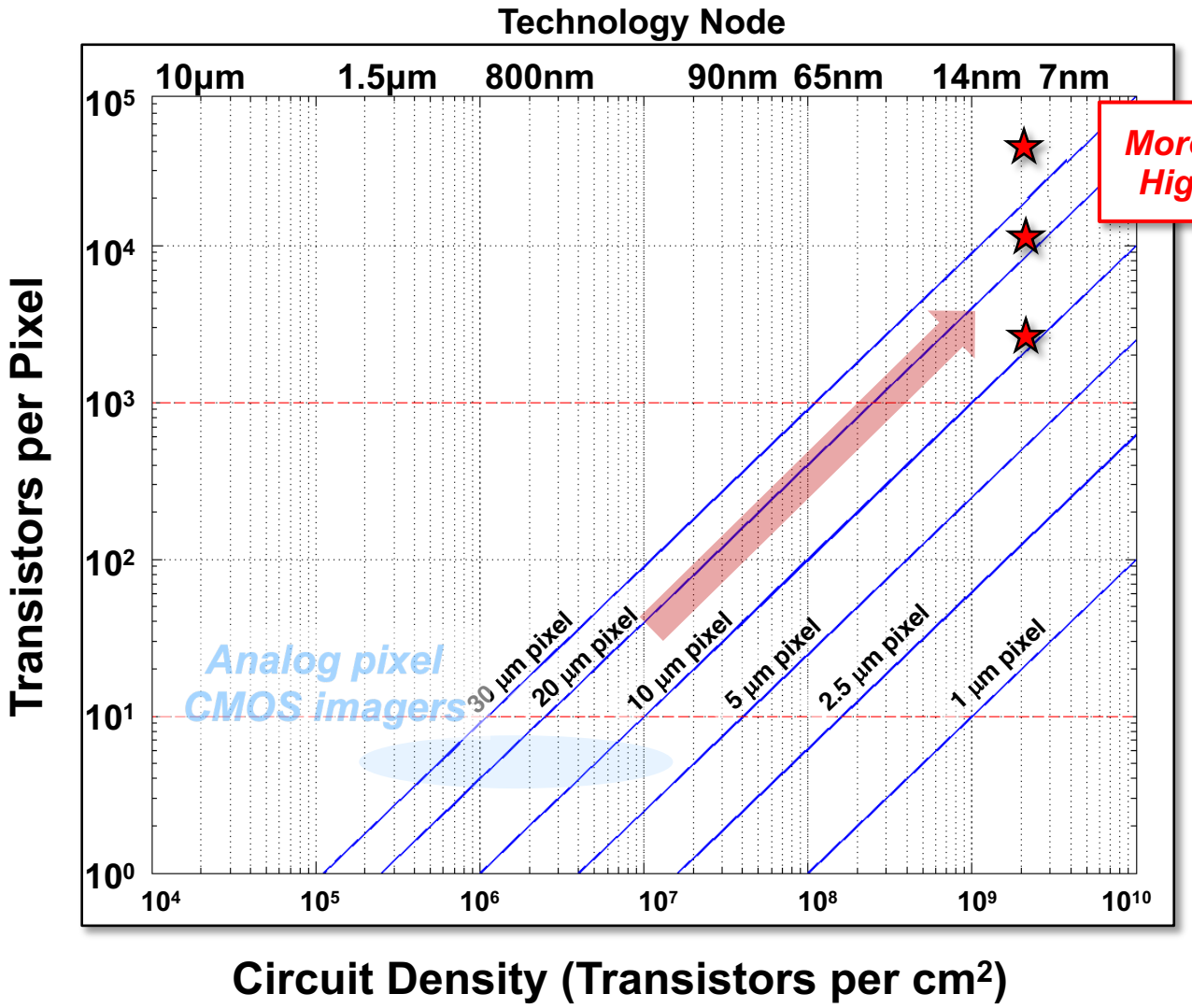
This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering or the United States Government.

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Evolution of High Performance Imagers



ROIC Fabrication Technology Landscape



More Transistors High Capability

As density increases, so does cost

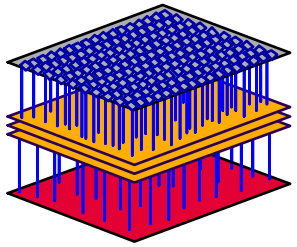
- **Mistakes are *expensive***
- **Only the biggest experiments can afford to fabricate**

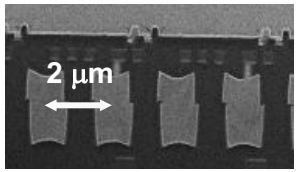
High Performance Imager Availability

- **How can we bring advanced processing ROICs to more applications?**

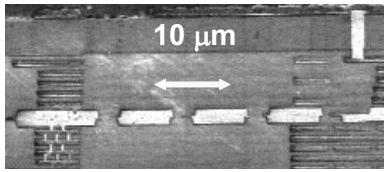
➢ **Increase hardware *modularity* and *reconfigurability* to reuse ROICs and amortize costs**

Hardware Modularity: Split the ROIC

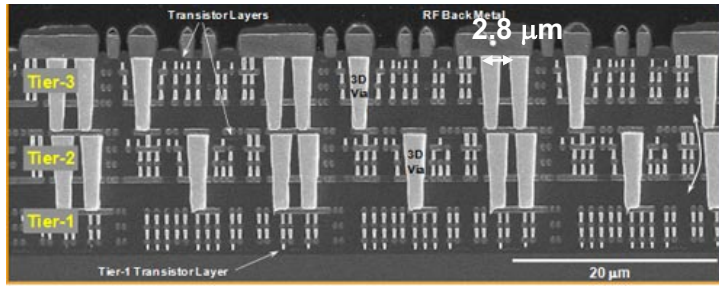




Direct Bond Interconnect (Ziptronix)

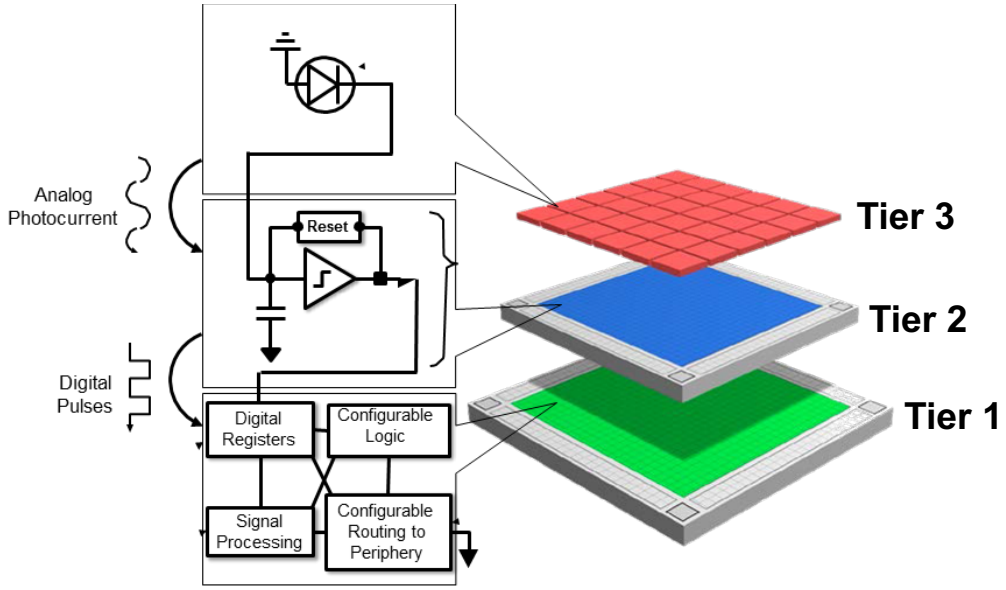


Cu-Cu bond (Tezzeron)



Through-oxide-vias (Lincoln Lab)

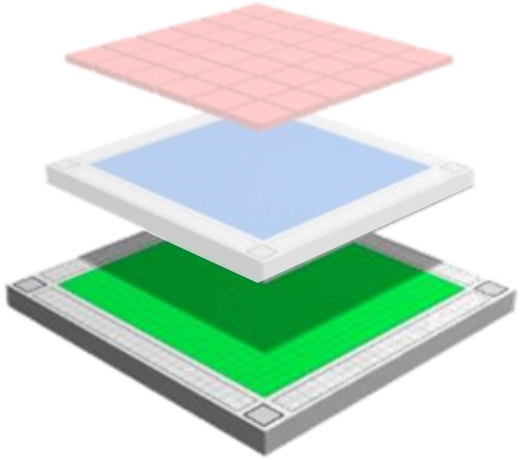
IO pitch < 10 μm
Short interconnect lengths



Tier 3	Waveband Specific Detector Array
Tier 2	Detector Specific Analog Interface
Tier 1	Reconfigurable Digital Circuit with Integrated FPGA Processing

Demonstrating Modularity and Reconfigurability

Two tier 1 chips fabricated under the DARPA ReImagine program



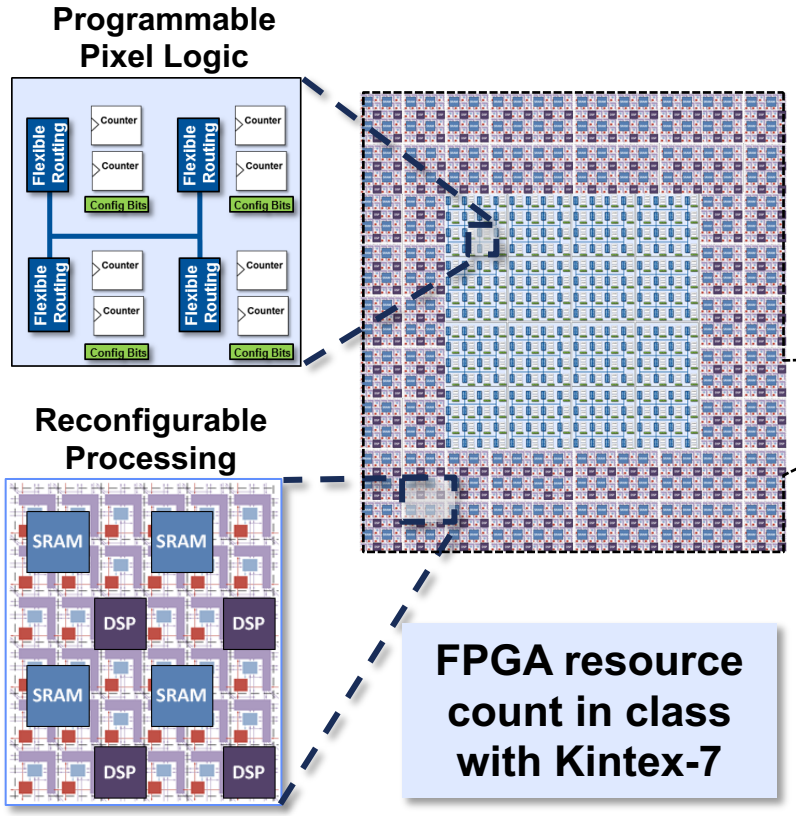
Tier 1

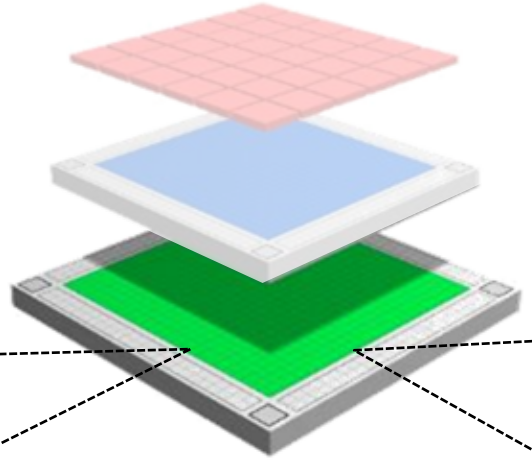
- **Advanced CMOS**
- **1280 × 1024**
- **12µm pitch**

Demonstrating Modularity and Reconfigurability

Griffin

FPGA-based architecture



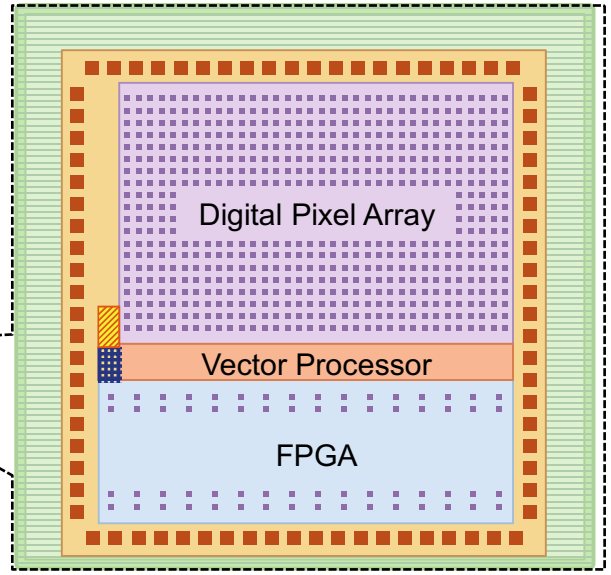


- Advanced CMOS
- 1280 × 1024
- 12µm pitch

FPGA resource count in class with Kintex-7

Manticore

SoC-based architecture



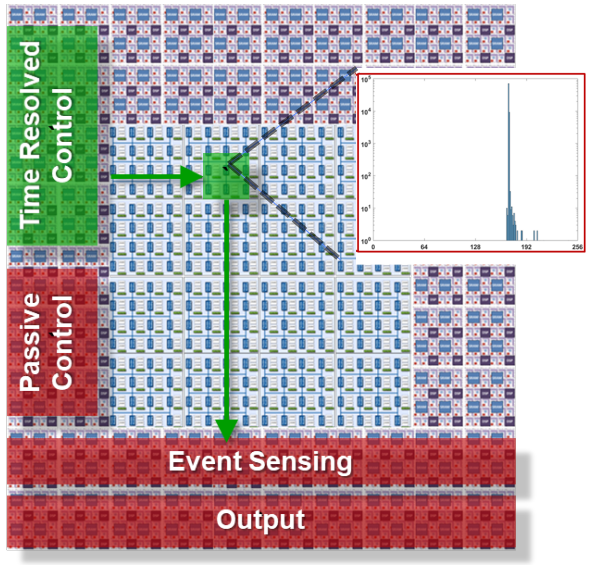
- Programmable pixel array processor
- SIMD vector processor
- FPGA
- Microprocessor control core

Each 16x16 macropixel has comparable MOPS to a Pentium I

Using Smart Imagers

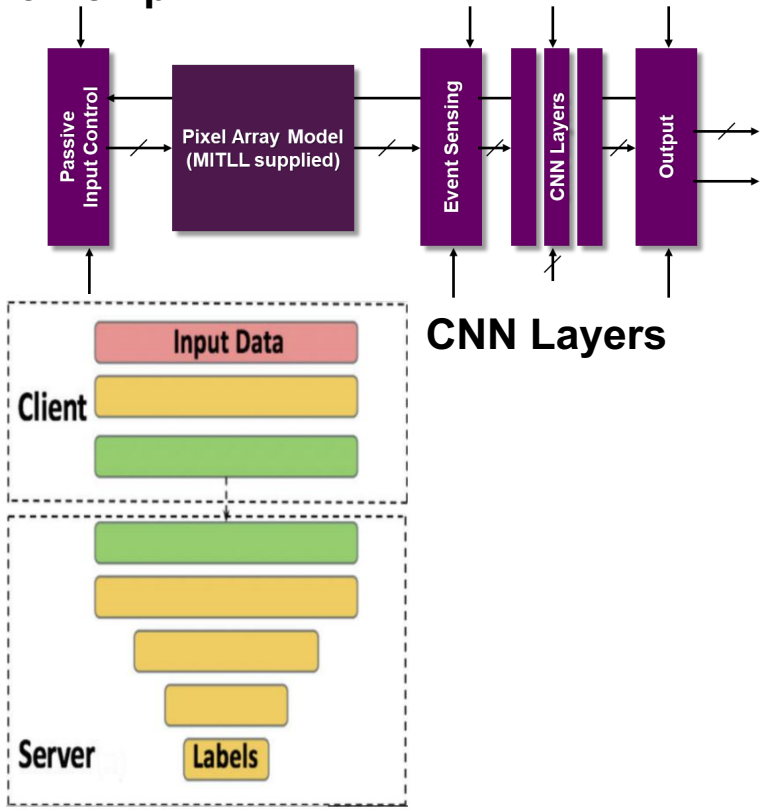
Multimodal Sensing

Self-trigger a change in sensing based on events



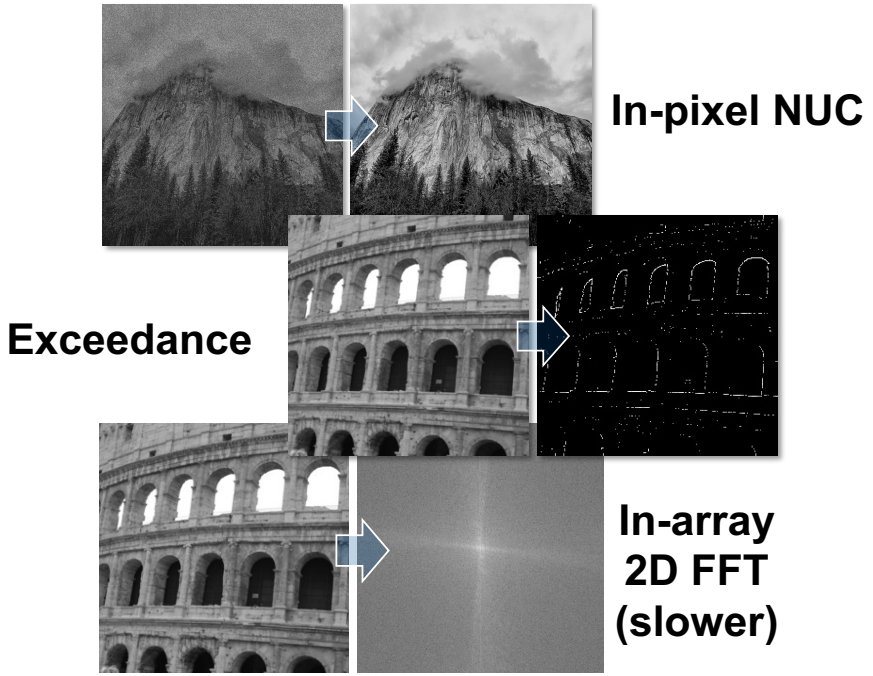
Machine Learning

CNNs are feasible to implement on-chip



Conventional Processing

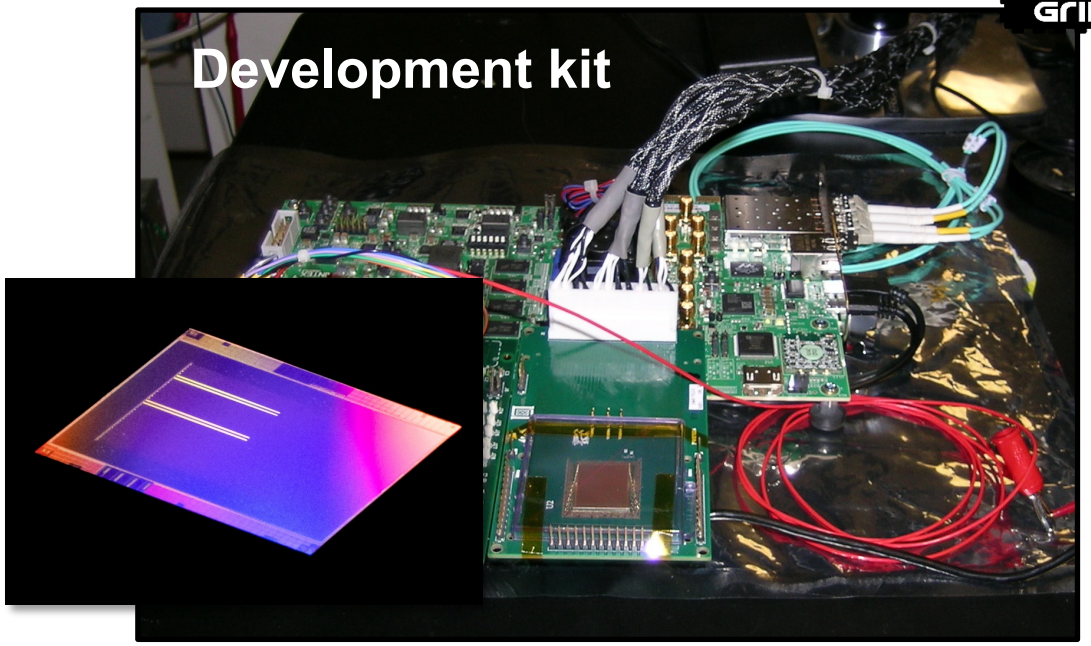
Many types of on-chip processing at rates in excess of 100,000 FPS



More sophisticated adaptive processing is possible in-pixel and in-array

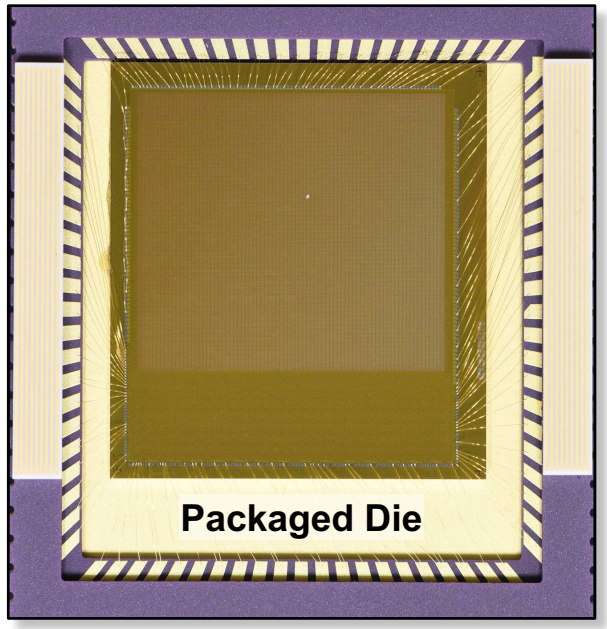
ReImagine ROIC Status

Griffin



- Multiple partners with Griffin tier-2 chips
- Griffin T1-only hardware development kit shipped January 2021

Manticore



- Multiple partners in process of tier 2 design
- Anticipate at least one bonded assembly to begin late 2021
- Demo camera in 2022

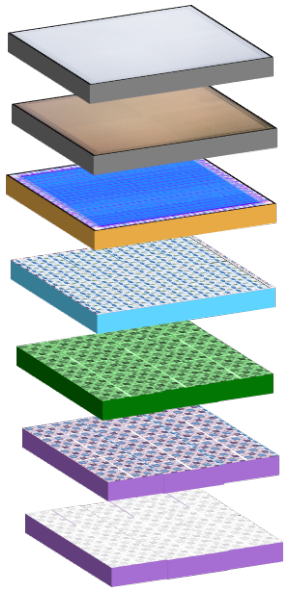
What's next?

Near Future

- Demonstration cameras
- Design refinement
- Open to collaborations on Griffin & Manticore
- Open to studying needs for science & space applications

Looking Ahead

- Is on-chip ML the path for perceptive imagers?
- Create deep stacking architectures for full modularity
- Abuttability



Thank You

younger@ll.mit.edu

MIT/LL ReImagine Team

Rich Younger
Valerie Finnemeyer
Jon Frechette
Sue Burzyk
Brian Tyrrell

Matt Stamplis
Jonathan Leu
Peter Grossmann

Matt Gregory
George Jordy
Tom Cheng
Jim Wey
Alice Lee

Ana-Maria Mandrila Vacca
Brian Yu

Maria Blood
Glenn Garvey
Kate Gillis
Philemon Chose

Kate Thurmer
Jordan Lahanas
Tom Ross
Ian Brown

Tony Kryzak
Austin Holloway
Tony Soares
David Volfson
Jerry Lipson
Tom Ferguson
Mike Cooper

Renee Lambert
Donna Yost
Sandra Lee

Hannah Whisnant
Praneeth Vepakomma (MIT)
Tom Karolyshyn
Greg Rowe

Farhan Adil
Phillip Bailey
Hernan Castro
Dan Santiago
Domenic Terranova

Tim Smith
Dmitriy Kaplan
Bob D'Ambra

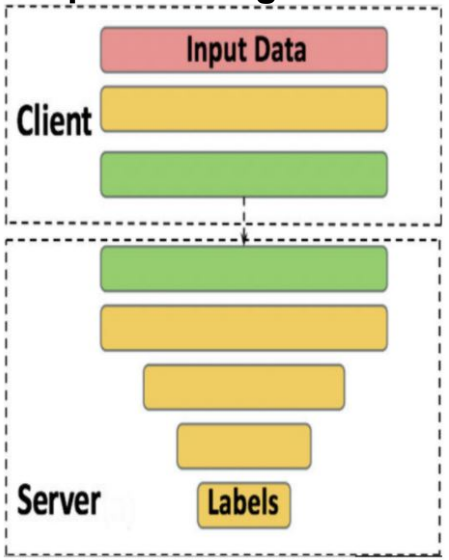
Tim Gagnon
Erik Duerr
Eric Dauler
Dan Ripin
Kaitlyn Dixon
Stephan Chase
Kenny Sims
Elaine Swenson

Designing and producing these chips is the product of a large team of professionals

Machine Learning Feasibility

- **Character recognition example using client/server split learning architecture**
  - **Split learning separates input data from full body of inference network**
- **Used hls4ml to translate character recognition split learning model in Keras to Vivado HLS**
- **Results: This model should, with further optimizations, fit on Griffin**
  - **Need to investigate network compression (pruning) & DSP re-use**
  - **2D convolutional layers were a challenge for hls4ml (as of v0.1.6)**

Simple split learning architecture



Character Recognition Model FPGA Resource Utilization

Resource	Split Learning Model	Kintex UltraScale 115	Griffin FPIA
LUTs	92,395	663,390	275,032
I/O	634*	702	275
DSP tiles	3,286	5,530	1,736
BRAM	60.5	2,160	820

* Includes image

Whisnant, H. K. (2020). *Split learning on FPGAs* (Master's Thesis, Massachusetts Institute of Technology).
Duarte *et al.*, "Fast inference of deep neural networks in FPGAs for particle physics", JINST 13 P07027 (2018), arXiv:1804.06913.