AI Chiplet Set with Supply Chain Security

Paul Franzon

Cirrus Logic Distinguished Professor,

Director of Graduate Programs

Department of Electrical and Computer Engineering,

NC State University

paulf@ncsu.edu

Chiplet Summit, Jan, 2023

Outline

- AI chiplet set – 2.5D
  - RISC V + accelerators
- AI chiplet – 3D pnm
  - 3D logic on memory
- Embedded RFID to secure supply chain
- CAD tools for Chiplets

AI chiplet set

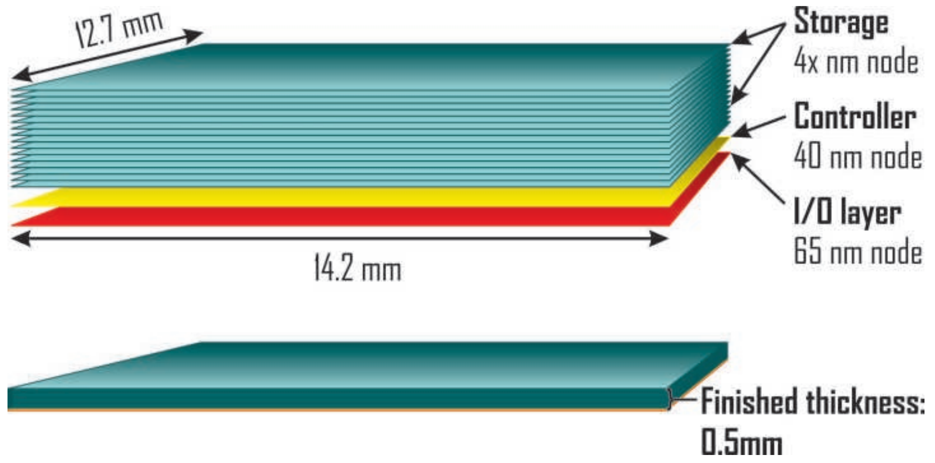
- Unpublished material – please contact Franzon if you are interested

Outline

- AI chiplet set (1) – 2.5D
  - RISC V + accelerators
- AI chiplet – 3D PNM
- 3D logic on memory
- Embedded RFID to secure supply chain
- CAD tools for Chiplets

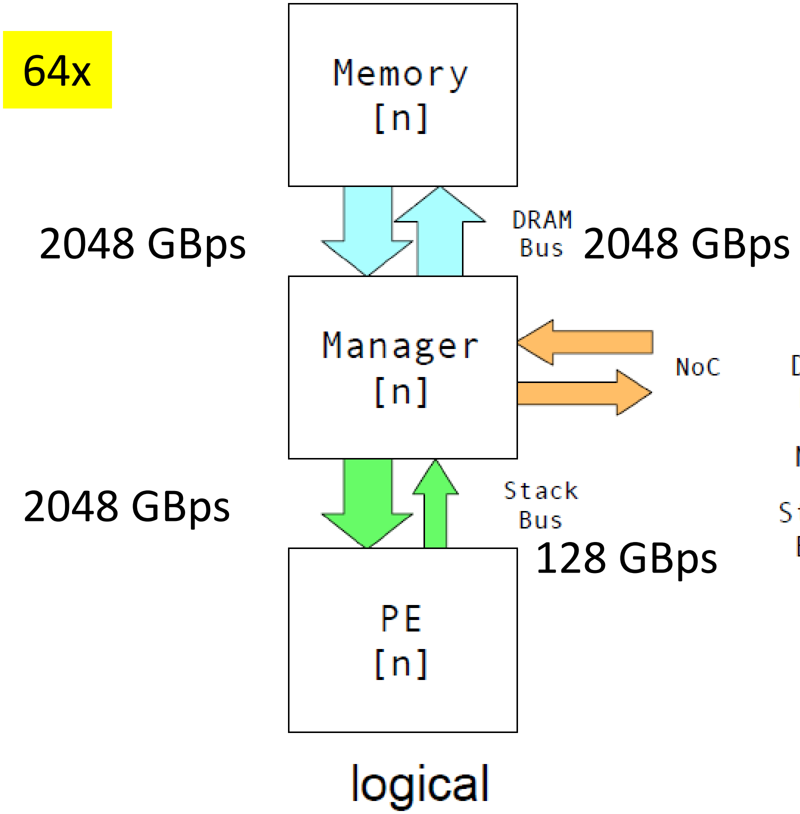
Logic on DRAM CNN Accelerator

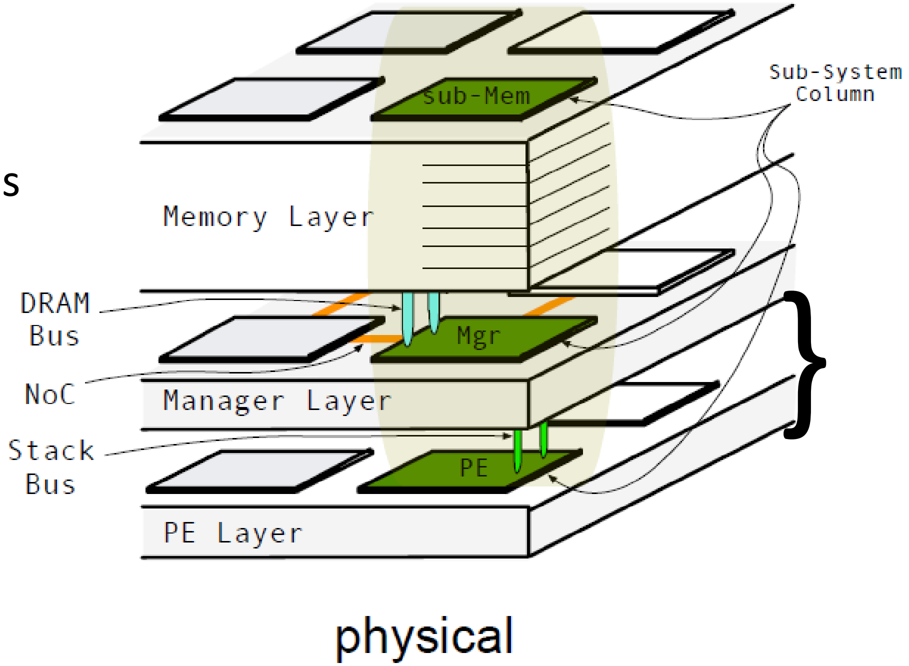
- Neural Networks getting very large
  - GPT3 - 175 B parameters (570 GB)
- Need lots of capacity and lots of bandwidth
  - Solution : Custom DRAM
- Design exercise based on modified Tezzaron 64 Gb DiRAM4



Using Customized DiRAM4

- Inference engine:

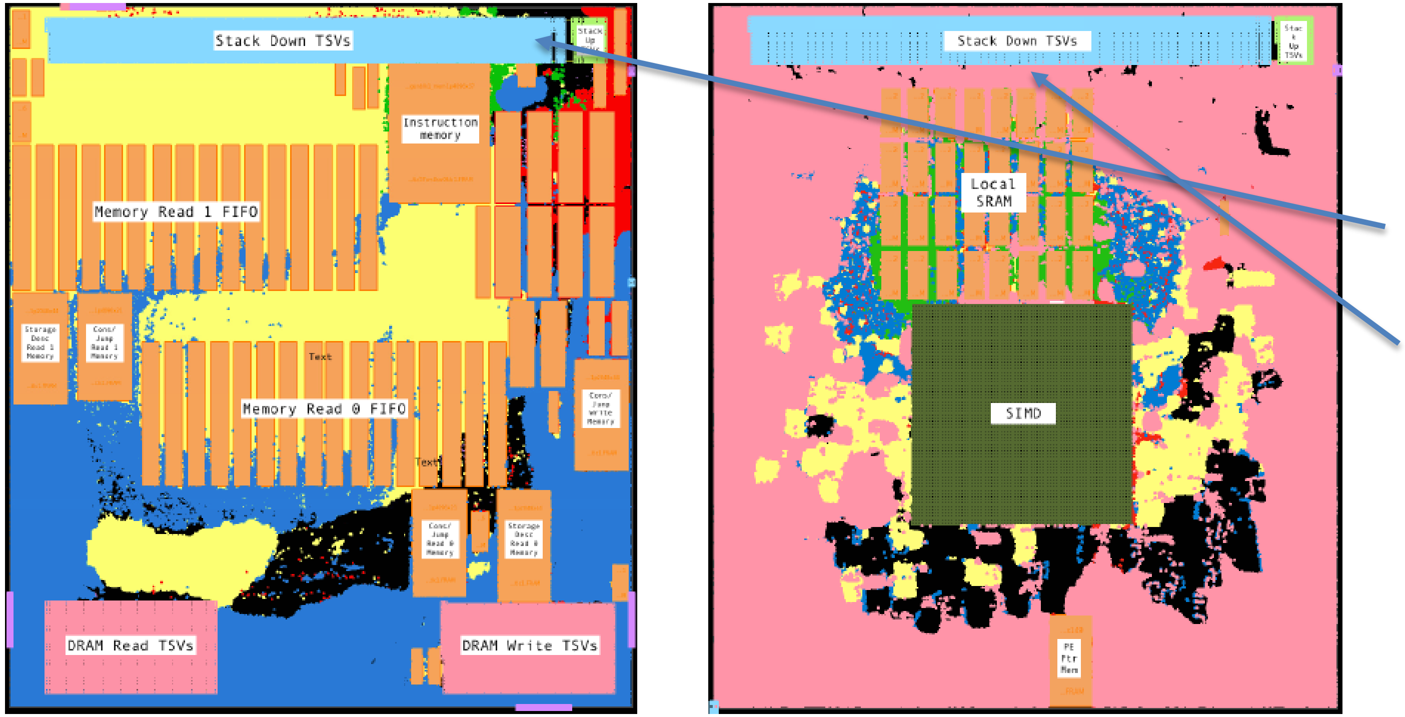




131 Tbps peak total bandwidth

Streaming Design : PE

- 65 nm CMOS : 2.4 x 2.7 mm (fits DRAM bank stack)
- Weight +data storage in DRAM



Management Chip

Processing Element

5 um TSV pitch

4200 TSVs

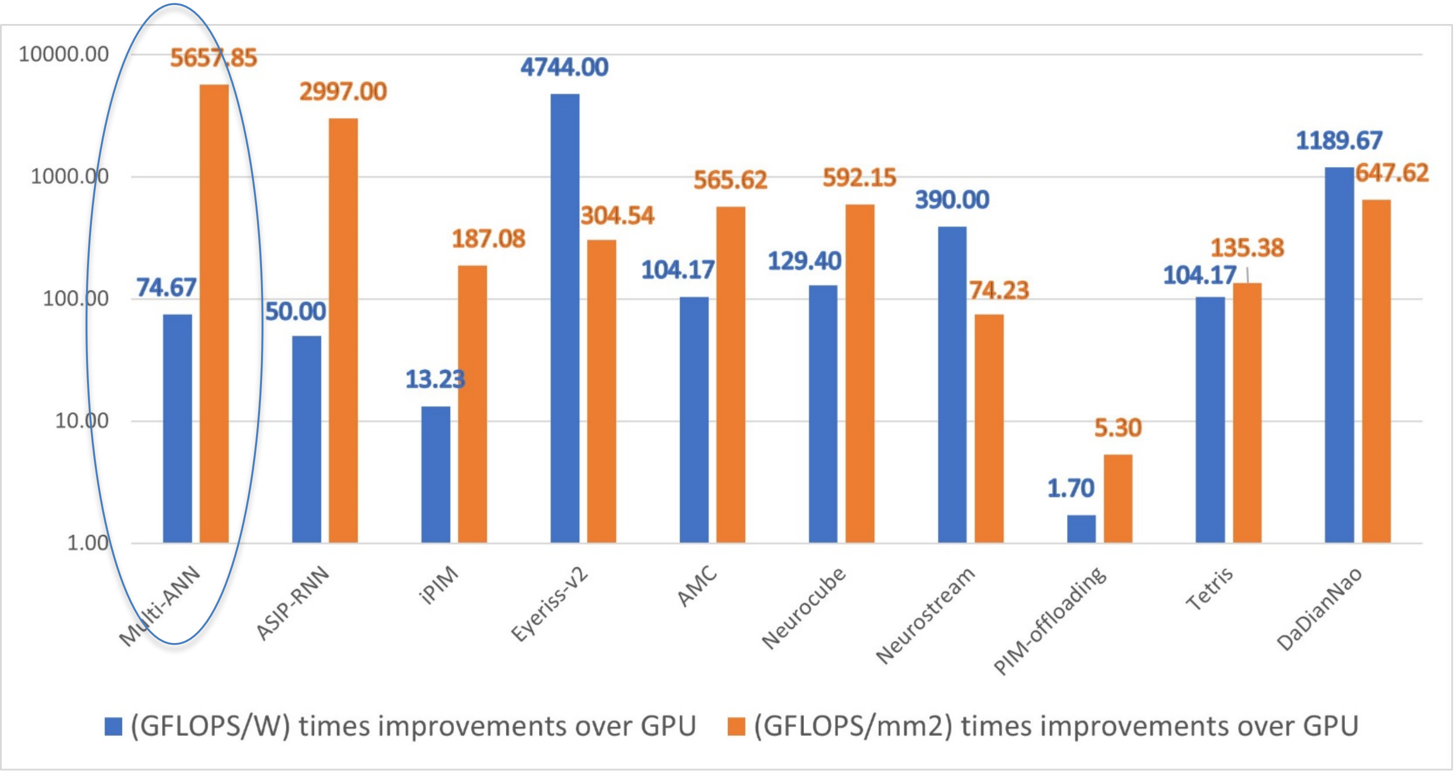
2500 TSVs

Baker, S3S, 2018

Baker, 3DIC 2021

Analysis – Performance Efficiencies

- Gain in performance efficiencies of different MNC architectures over the GPU baseline is computed.

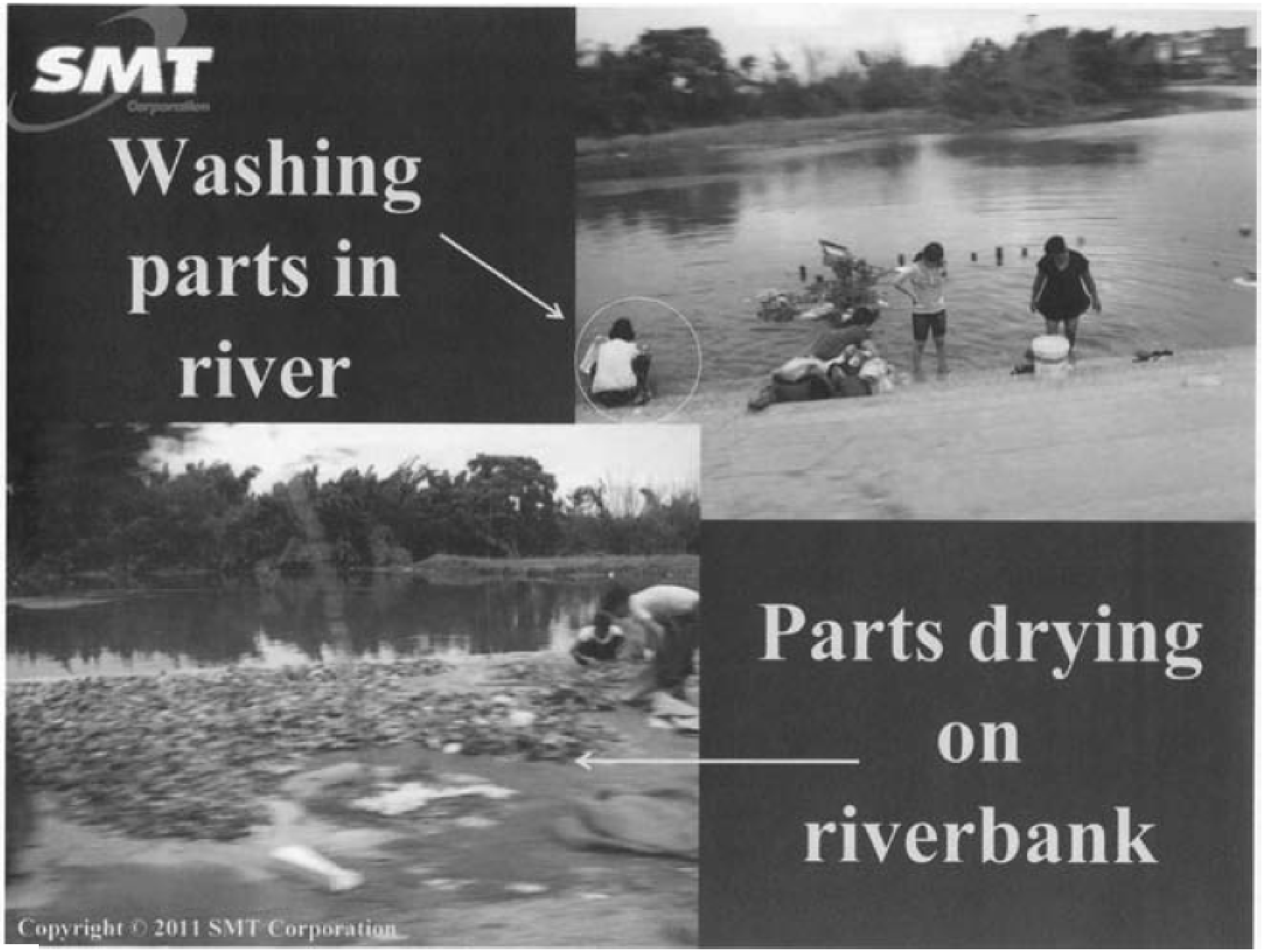


Ravichandran, 3DIC 2021

Outline

- AI chiplet set – 2.5D
  - RISC V + accelerators
- AI chiplet – 3D PNM
  - 3D logic on memory
- Embedded RFID to secure supply chain
- CAD tools for Chiplets

Securing the Semiconductor Supply Chain



Washing parts in river

Parts drying on riverbank

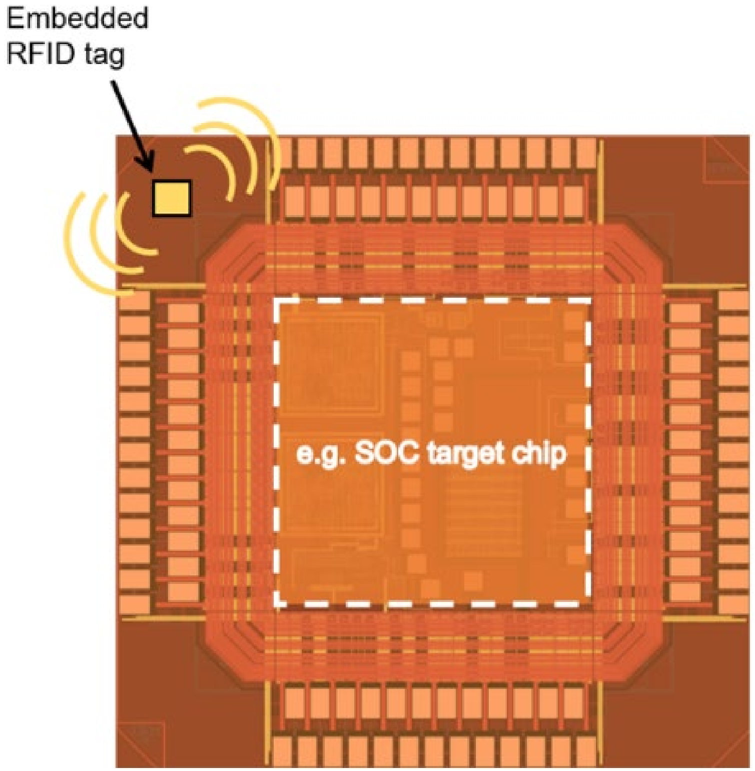
Copyright © 2011 SMT Corporation

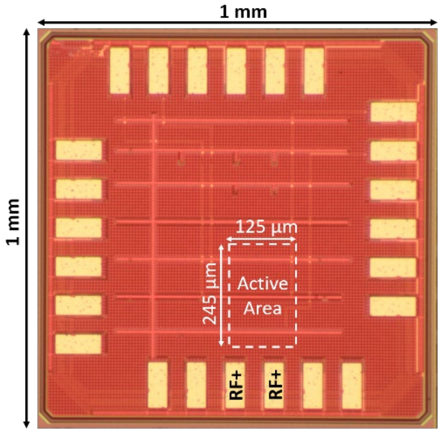
S. Hrg. 112–340

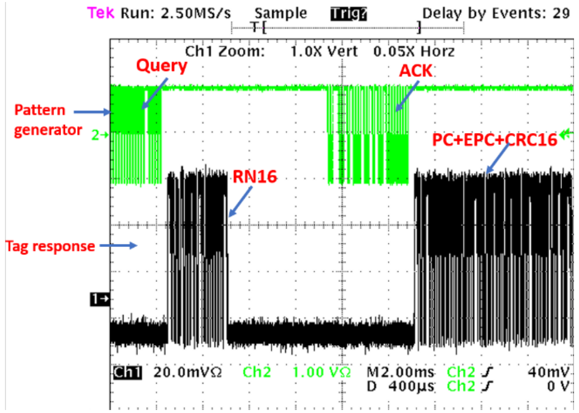
THE COMMITTEE'S INVESTIGATION INTO COUNTERFEIT ELECTRONIC PARTS IN THE DEPARTMENT OF DEFENSE SUPPLY CHAIN

Embedded RFID

- Mostly digital tiny embedded RFID – wirelessly interrogated to verify authenticity
- Chip cryptographically recognized
- Tracked via secure database
- Chip does not need to be powered



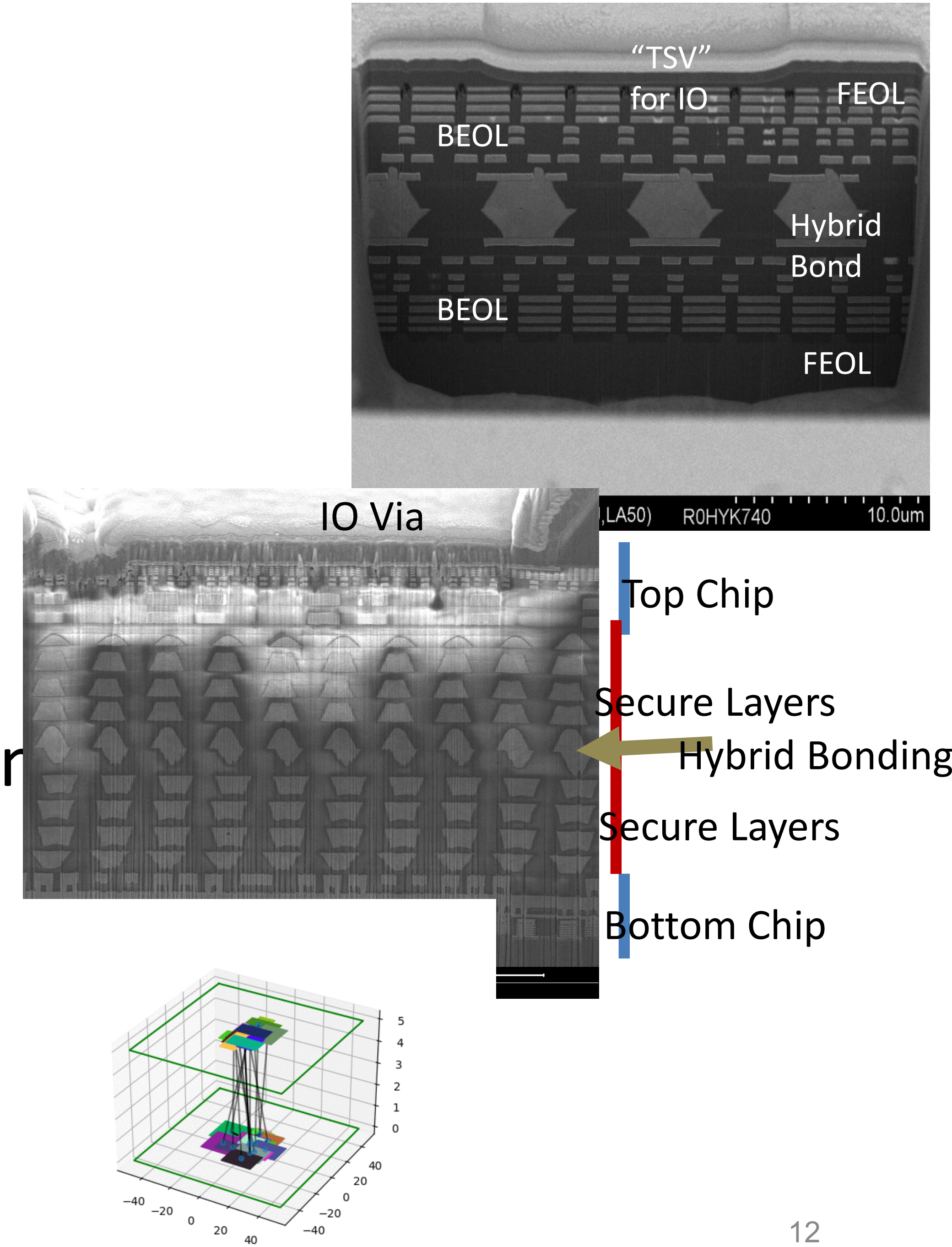




55 nm prototype

Outline

- AI chiplet set – 2.5D
  - RISC V + accelerators
- AI chiplet – 3D PNM
  - 3D logic on memory
- Embedded RFID to secure supply chair
- CAD tools for Chiplets
  - Performance based partitioning
  - Obfuscation based partitioning
  - 2.5D and 3D floorplanning



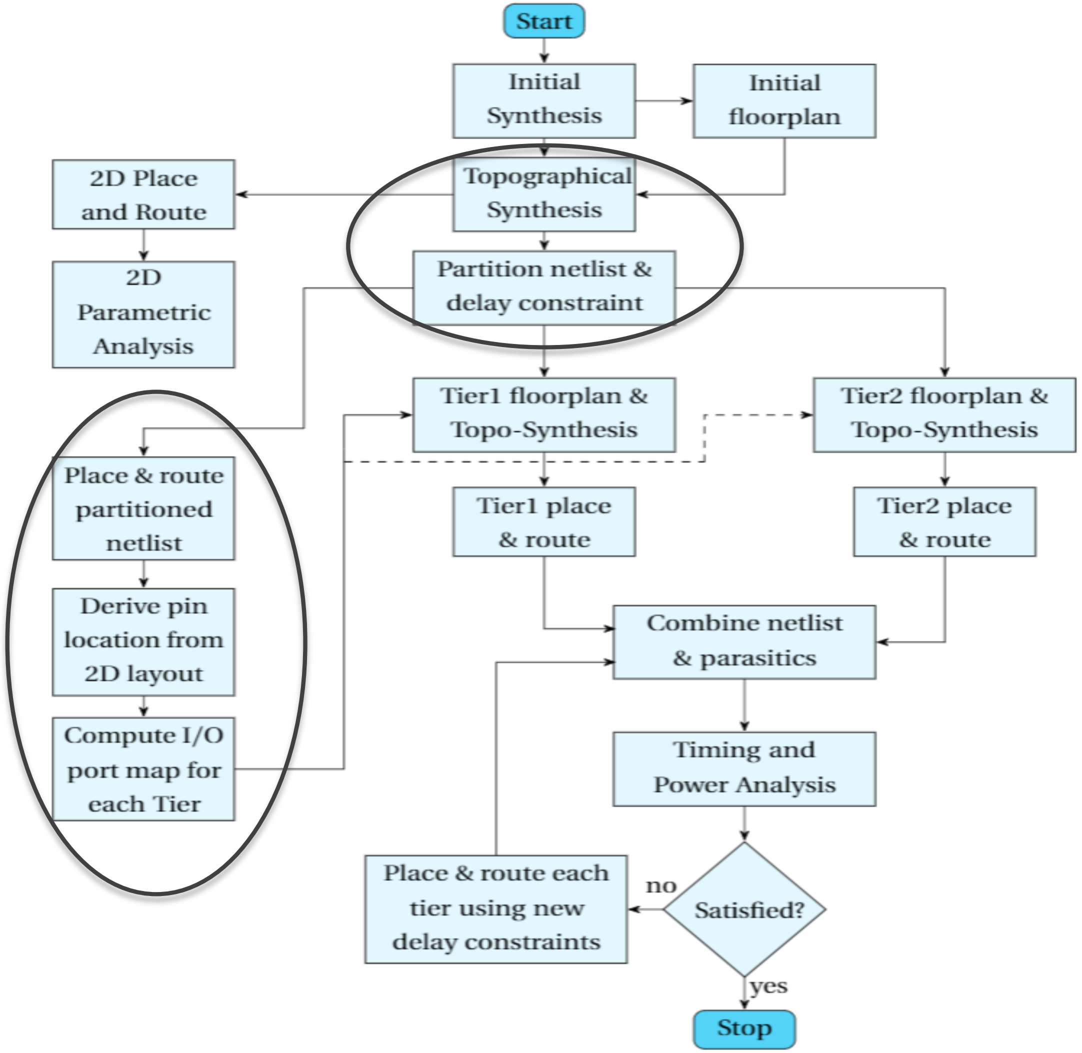
Conclusions

- Scalable AI chiplet set for edge inference
- 130 Tbps DRAM for core inference
- Embedded RFID for securing supply chain
- CAD flows for
  - Performance based partitioning
  - Obfuscation based partitioning
  - 2.5D and 3D floorplanning

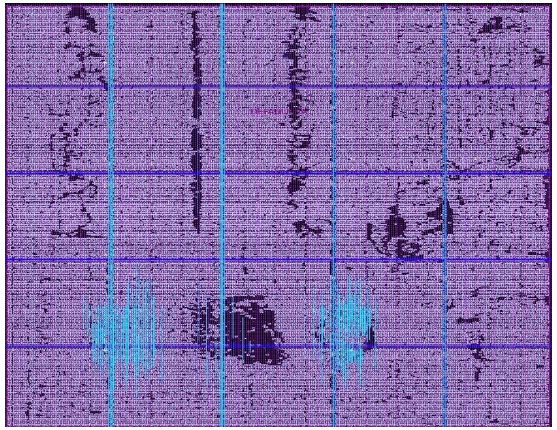
Acknowledgements

Faculty: Rhett Davis, Michael B. Steer, Eric Rotenberg, James Tuck, Huiyang Zhou
Professionals: Steven Lipa, Eric Wyers
Students: Brandon Dwiel, Joshua Ledford, Jong Beom Park, Lee Baker, Sumon Dey, Weifu Li, Josh Stevens, Teddie Nigussie, Tse-Han Pan, W. Shep Pitts, Joshua Stevens, Han Pan, Shafin Amin,
Hua Hao, Samson Melamed, Peter Gadfort, Akalu Lentiro, Zhou Wang, Marcus Tishibanqu, Elliott Forbes, Randy Wilkiansano, Joonmu Hu, Prasanth Ravichandra,
Shivam Priyadarshi, Christopher Mineo, Julie Oh, Won Ha Choi, Ambrish Sule, Gary Charles, Thor Thorolfsson,
Department of Electrical and Computer Engineering
NC State University

3D Partitioning using Hybrid Bonding


Start
Initial Synthesis
Initial floorplan
2D Place and Route
Topographical Synthesis
2D Parametric Analysis
Partition netlist & delay constraint
Tier1 floorplan & Topo-Synthesis
Tier2 floorplan & Topo-Synthesis
Place & route partitioned netlist
Tier1 place & route
Tier2 place & route
Derive pin location from 2D layout
Combine netlist & parasitics
Compute I/O port map for each Tier
Timing and Power Analysis
Place & route each tier using new delay constraints
no
Satisfied?
yes
Stop

2D Design – starting point for partitioning

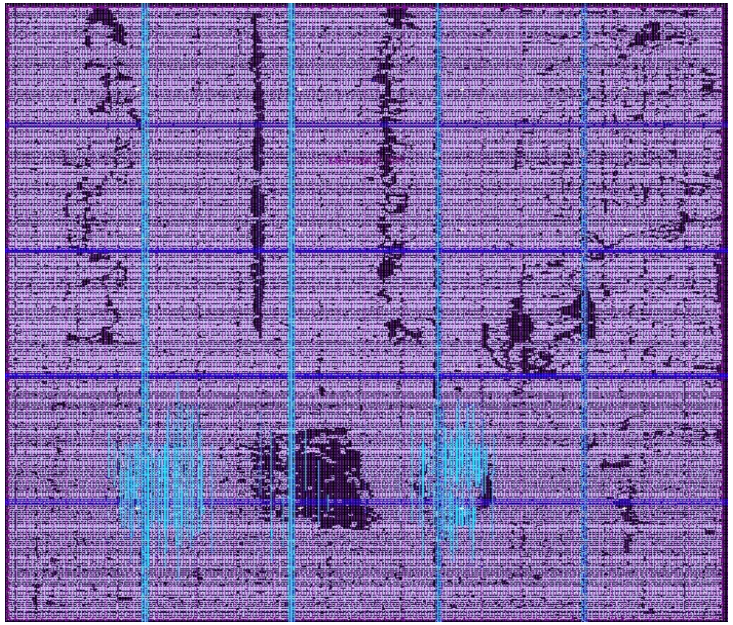


Goal : 50:50 area split

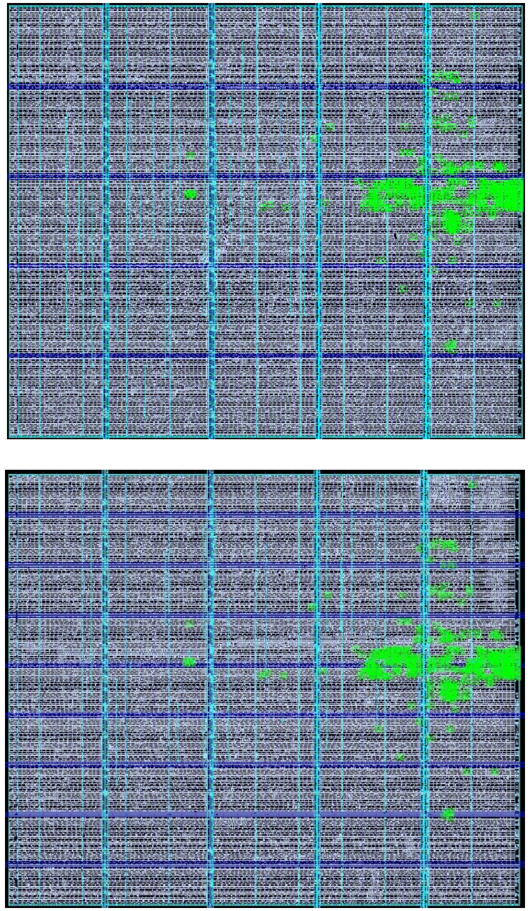
Partition nets based on net length, net power, achieving a 50:50 split and routability.

IO port map derived from 2D design

FFT 2D vs. 3D



FFT 2D (440 x 440 um)



FFT 3D (295 x 295 um)

FFT 2D vs 3D

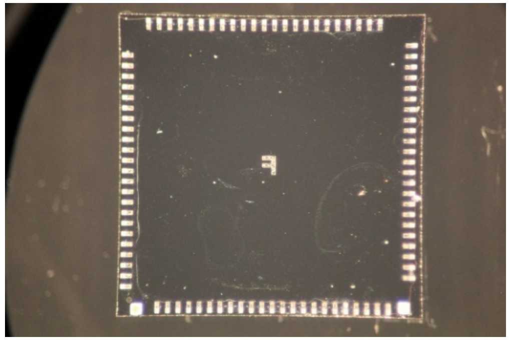
- 3D Design 3 : 9 metal stack, clock gating turn on, 5 ns clock

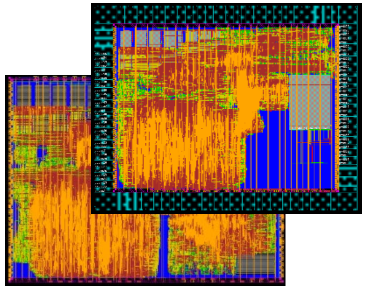
Parameter	2D	3D	Improvement
Clock Period (ns)	4.98	4.98	0%
Power (mW)	26.1	20.4	21.8%
Silicon area	193,600	174,050	11%

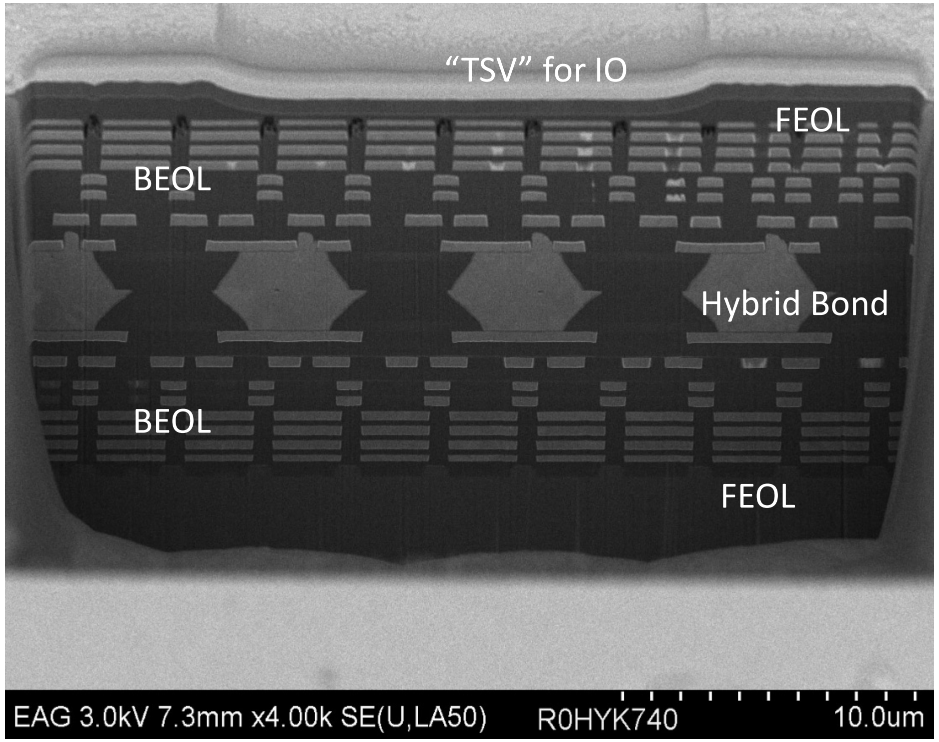
Cost saving

  - Alternatively, remove 3 metals from stack
  - 57% reduction in routed wire length

Heterogeneous Computing

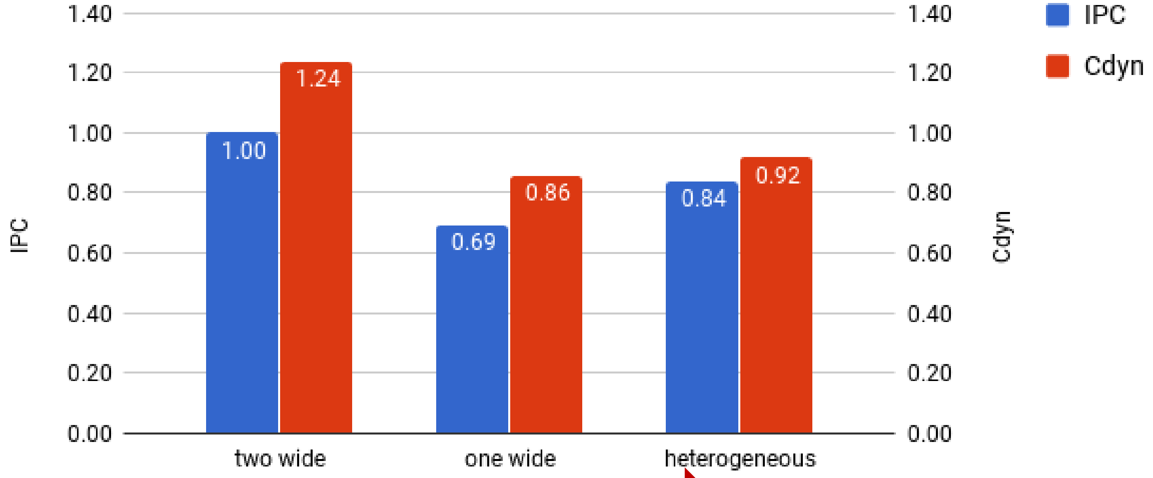






Heterogeneous Microprocessor

- Stack low-power and high-performance processors
- Swap threads vertically when workload warrants

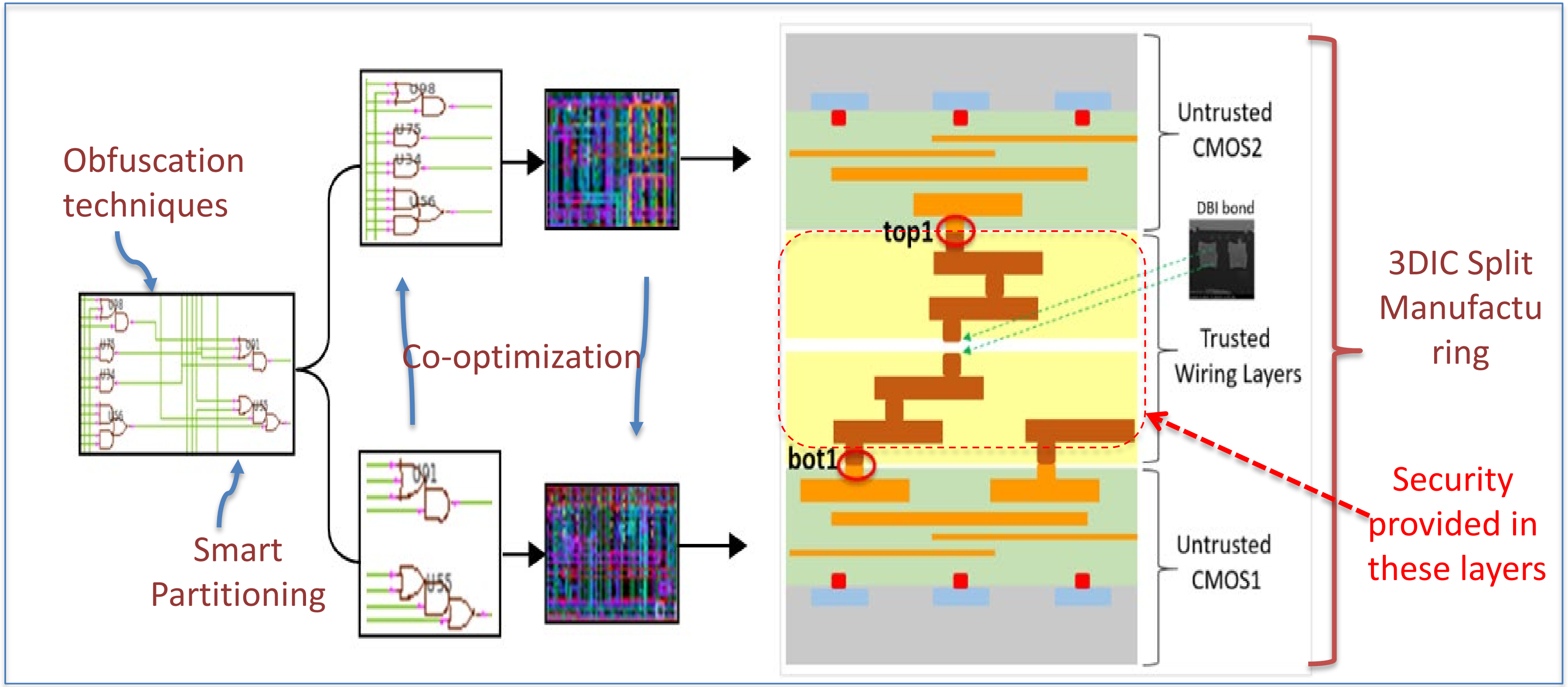


13% better performance/power

14% better performance/power

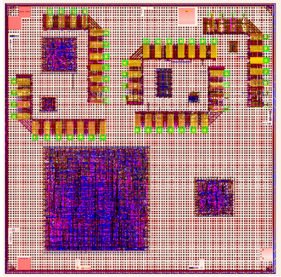
Split Fabrication for Design Obfuscation

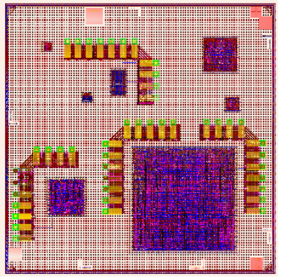
- Objective: Prevent reverse engineering of key design intent if CMOS layers are compromised

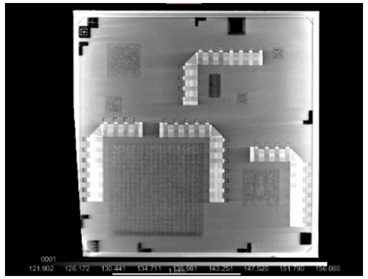


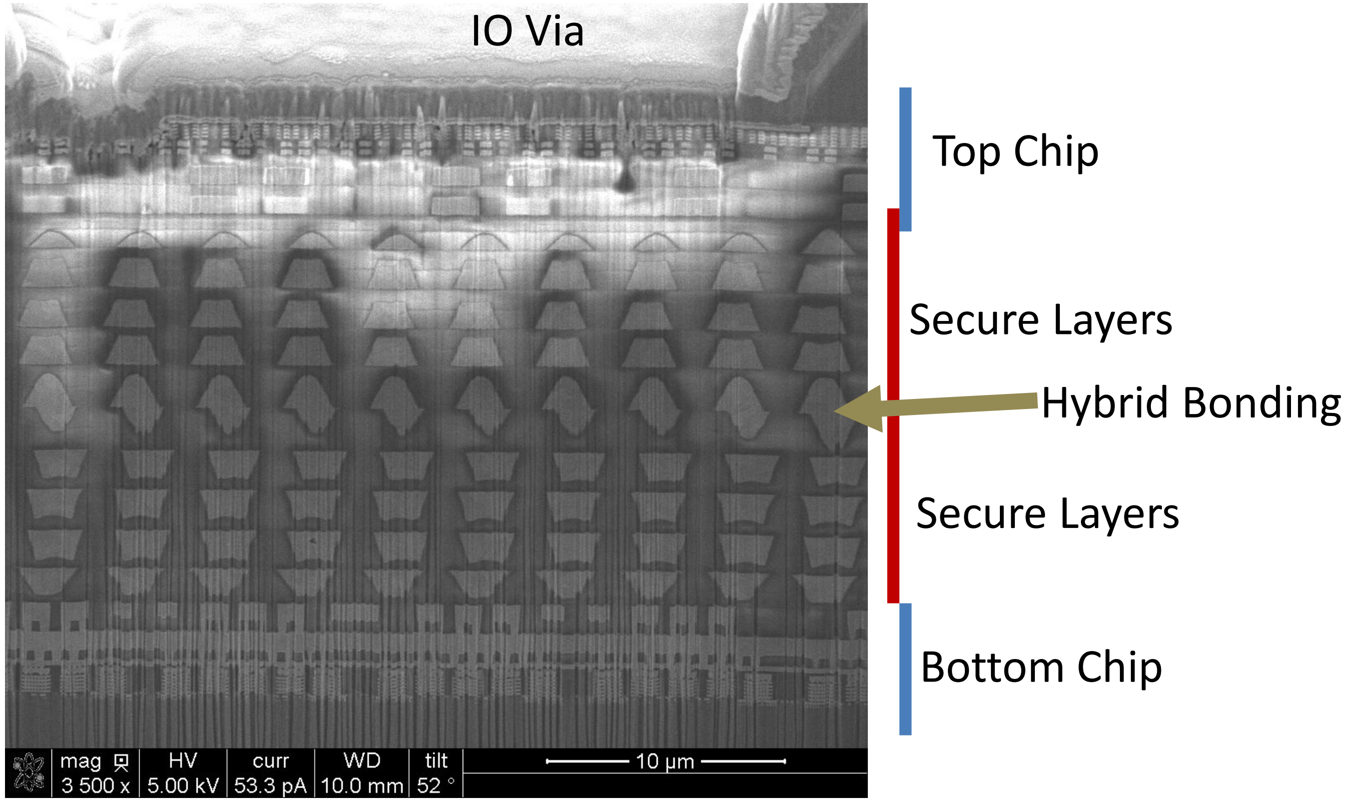
Nigussie, Franzon, 2021

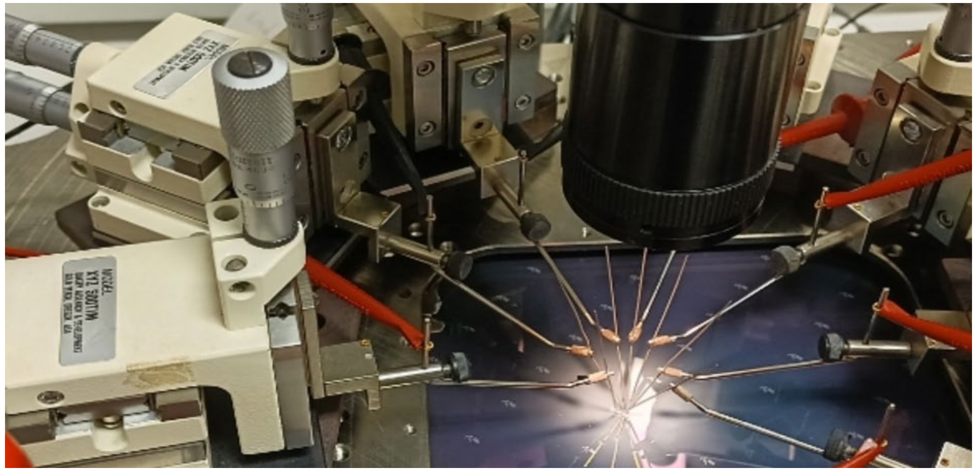
Split Fabrication

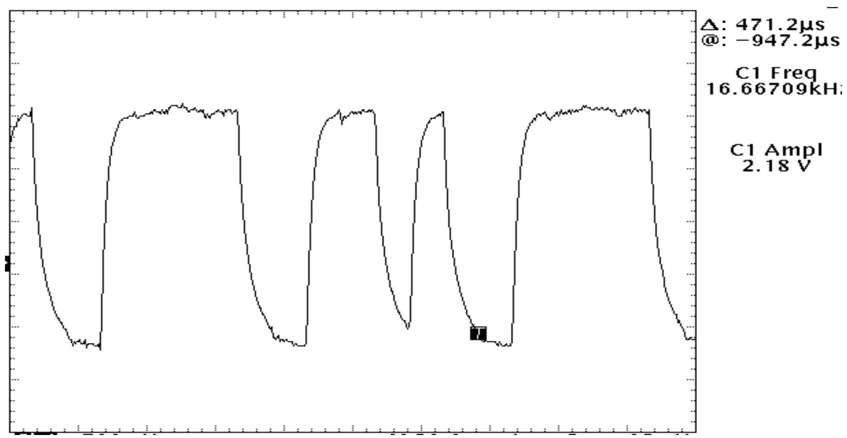








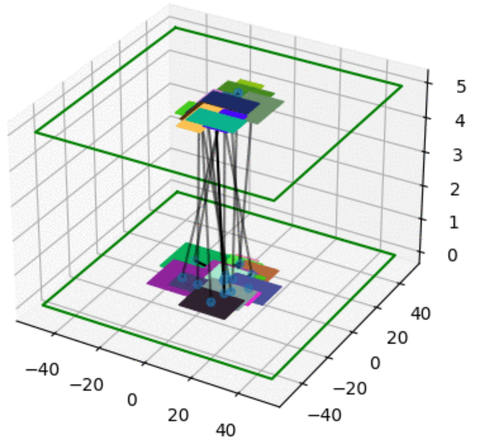


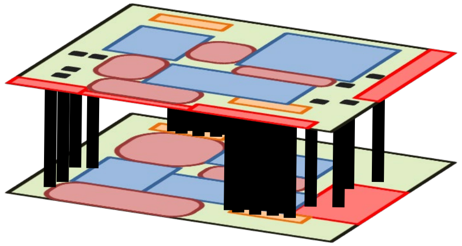


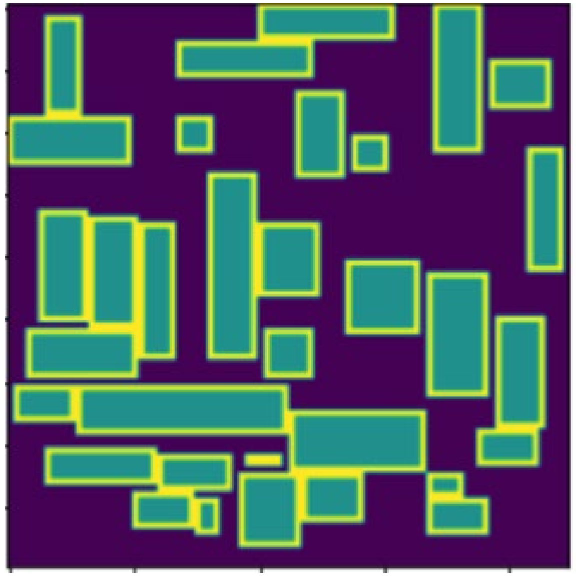
Nigussie, Franzon, 2021

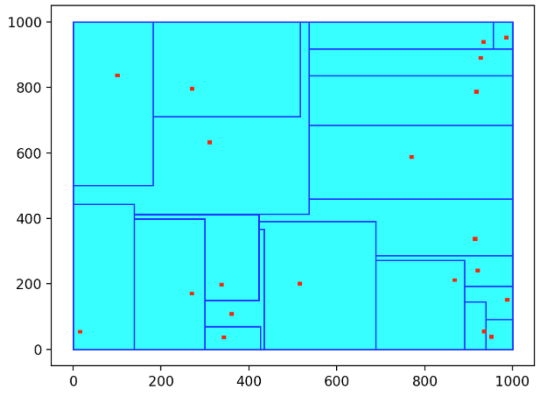
2.5D and 3D Floorplanning

- Classical and ML based approaches









Conclusions

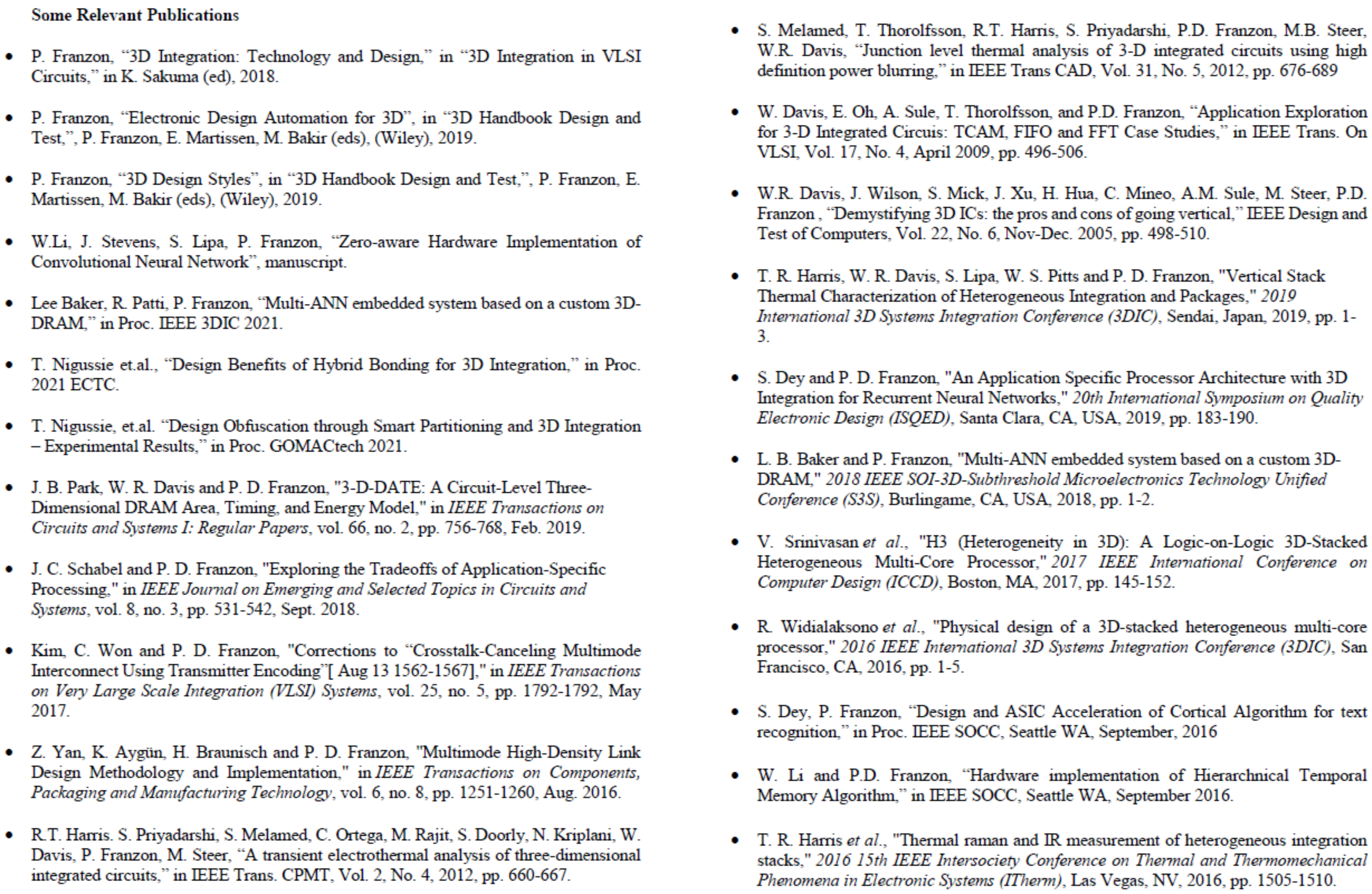
- Scalable AI chiplet set for edge inference
- 130 Tbps DRAM for core inference
- Embedded RFID for securing supply chain
- CAD flows for
  - Performance based partitioning
  - Obfuscation based partitioning
  - 2.5D and 3D floorplanning

22

Acknowledgements

Faculty: Rhett Davis, Michael B. Steer, Eric Rotenberg, James Tuck, Huiyang Zhou
Professionals: Steven Lipa, Eric Wyers
Students: Brandon Dwiel, Joshua Ledford, Jong Beom Park, Lee Baker, Sumon Dey, Weifu Li, Josh Stevens, Teddie Nigussie, Tse-Han Pan, W. Shep Pitts, Joshua Stevens, Han Pan, Shafin Amin,
Hua Hao, Samson Melamed, Peter Gadfort, Akalu Lentiro, Zhou Wang, Marcus Tishibanqu, Elliott Forbes, Randy Wilkiansano, Joonmu Hu, Prasanth Ravichandra,
Shivam Priyadarshi, Christopher Mineo, Julie Oh, Won Ha Choi, Ambrish Sule, Gary Charles, Thor Thorolfsson,
Department of Electrical and Computer Engineering
NC State University

Some of my references

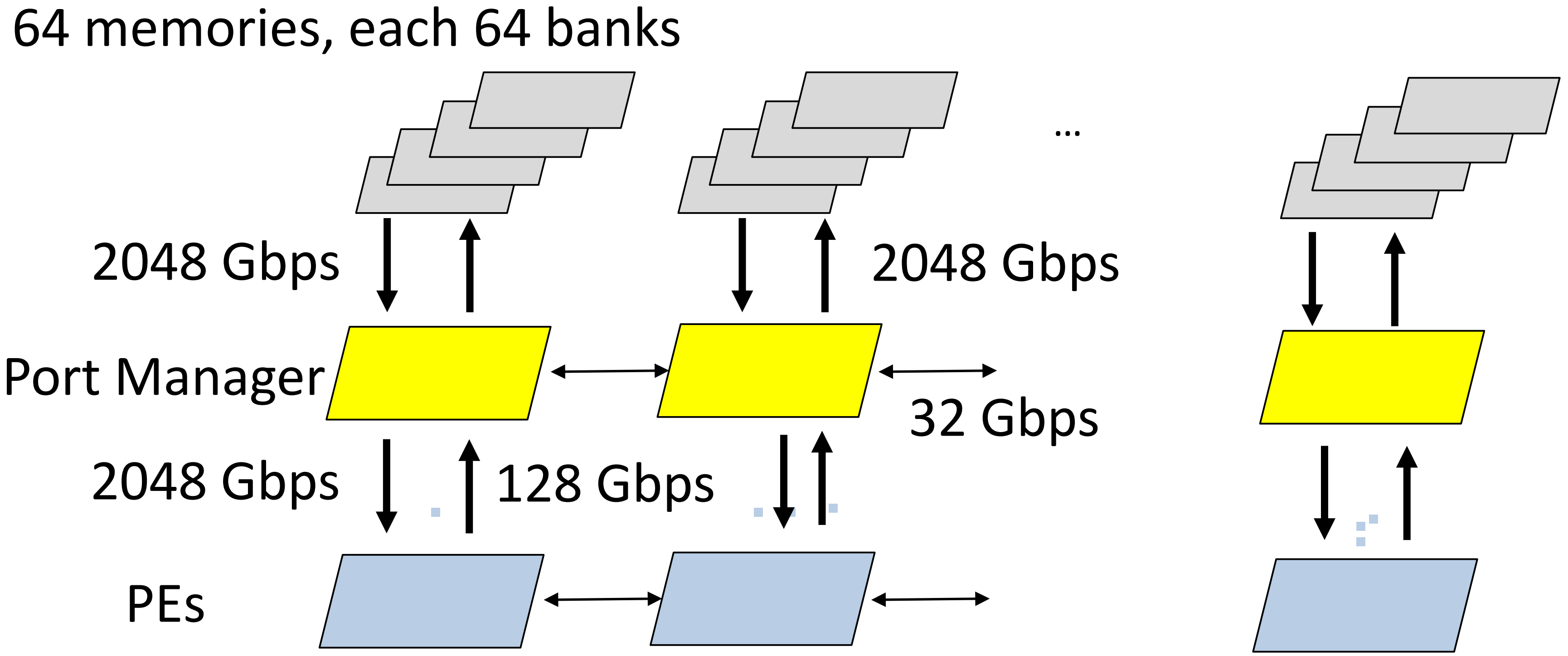


Some Relevant Publications

- P. Franzon, "3D Integration: Technology and Design," in "3D Integration in VLSI Circuits," in K. Sakuma (ed), 2018.
- P. Franzon, "Electronic Design Automation for 3D", in "3D Handbook Design and Test,", P. Franzon, E. Martissen, M. Bakir (eds), (Wiley), 2019.
- P. Franzon, "3D Design Styles", in "3D Handbook Design and Test,", P. Franzon, E. Martissen, M. Bakir (eds), (Wiley), 2019.
- W.Li, J. Stevens, S. Lipa, P. Franzon, "Zero-aware Hardware Implementation of Convolutional Neural Network", manuscript.
- Lee Baker, R. Patti, P. Franzon, "Multi-ANN embedded system based on a custom 3D-DRAM," in Proc. IEEE 3DIC 2021.
- T. Nigussie et.al., "Design Benefits of Hybrid Bonding for 3D Integration," in Proc. 2021 ECTC.
- T. Nigussie, et.al. "Design Obfuscation through Smart Partitioning and 3D Integration – Experimental Results," in Proc. GOMACtech 2021.
- J. B. Park, W. R. Davis and P. D. Franzon, "3-D-DATE: A Circuit-Level Three-Dimensional DRAM Area, Timing, and Energy Model," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 2, pp. 756-768, Feb. 2019.
- J. C. Schabel and P. D. Franzon, "Exploring the Tradeoffs of Application-Specific Processing," in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 3, pp. 531-542, Sept. 2018.
- Kim, C. Won and P. D. Franzon, "Corrections to "Crosstalk-Canceling Multimode Interconnect Using Transmitter Encoding"[Aug 13 1562-1567]," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 5, pp. 1792-1792, May 2017.
- Z. Yan, K. Aygün, H. Braunisch and P. D. Franzon, "Multimode High-Density Link Design Methodology and Implementation," in *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 6, no. 8, pp. 1251-1260, Aug. 2016.
- R.T. Harris. S. Priyadarshi, S. Melamed, C. Ortega, M. Rajit, S. Doorly, N. Kriplani, W. Davis, P. Franzon, M. Steer, "A transient electrothermal analysis of three-dimensional integrated circuits," in IEEE Trans. CPMT, Vol. 2, No. 4, 2012, pp. 660-667.
- S. Melamed, T. Thorolfsson, R.T. Harris, S. Priyadarshi, P.D. Franzon, M.B. Steer, W.R. Davis, "Junction level thermal analysis of 3-D integrated circuits using high definition power blurring," in IEEE Trans CAD, Vol. 31, No. 5, 2012, pp. 676-689
- W. Davis, E. Oh, A. Sule, T. Thorolfsson, and P.D. Franzon, "Application Exploration for 3-D Integrated Circuits: TCAM, FIFO and FFT Case Studies," in IEEE Trans. On VLSI, Vol. 17, No. 4, April 2009, pp. 496-506.
- W.R. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A.M. Sule, M. Steer, P.D. Franzon , "Demystifying 3D ICs: the pros and cons of going vertical," IEEE Design and Test of Computers, Vol. 22, No. 6, Nov-Dec. 2005, pp. 498-510.
- T. R. Harris, W. R. Davis, S. Lipa, W. S. Pitts and P. D. Franzon, "Vertical Stack Thermal Characterization of Heterogeneous Integration and Packages," *2019 International 3D Systems Integration Conference (3DIC)*, Sendai, Japan, 2019, pp. 1-3.
- S. Dey and P. D. Franzon, "An Application Specific Processor Architecture with 3D Integration for Recurrent Neural Networks," *20th International Symposium on Quality Electronic Design (ISQED)*, Santa Clara, CA, USA, 2019, pp. 183-190.
- L. B. Baker and P. Franzon, "Multi-ANN embedded system based on a custom 3D-DRAM," *2018 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, Burlingame, CA, USA, 2018, pp. 1-2.
- V. Srinivasan *et al.*, "H3 (Heterogeneity in 3D): A Logic-on-Logic 3D-Stacked Heterogeneous Multi-Core Processor," *2017 IEEE International Conference on Computer Design (ICCD)*, Boston, MA, 2017, pp. 145-152.
- R. Widialaksono *et al.*, "Physical design of a 3D-stacked heterogeneous multi-core processor," *2016 IEEE International 3D Systems Integration Conference (3DIC)*, San Francisco, CA, 2016, pp. 1-5.
- S. Dey, P. Franzon, "Design and ASIC Acceleration of Cortical Algorithm for text recognition," in Proc. IEEE SOCC, Seattle WA, September, 2016
- W. Li and P.D. Franzon, "Hardware implementation of Hierarchical Temporal Memory Algorithm," in IEEE SOCC, Seattle WA, September 2016.
- T. R. Harris *et al.*, "Thermal raman and IR measurement of heterogeneous integration stacks," *2016 15th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, Las Vegas, NV, 2016, pp. 1505-1510.

Modified DiRAM

- 130 Tbps of sustainable memory bandwidth



Fujitsu 55 nm

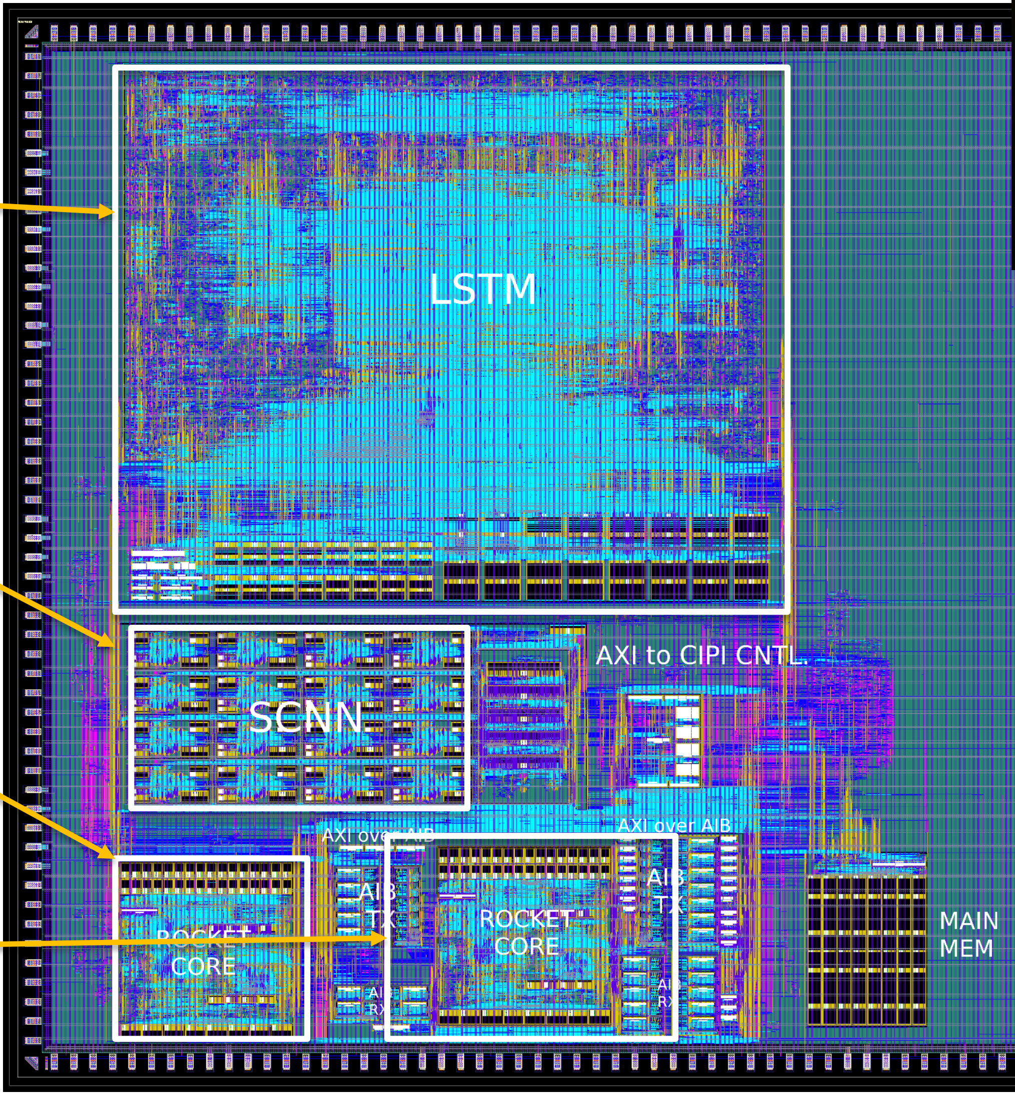
Simulates at 250 MHz

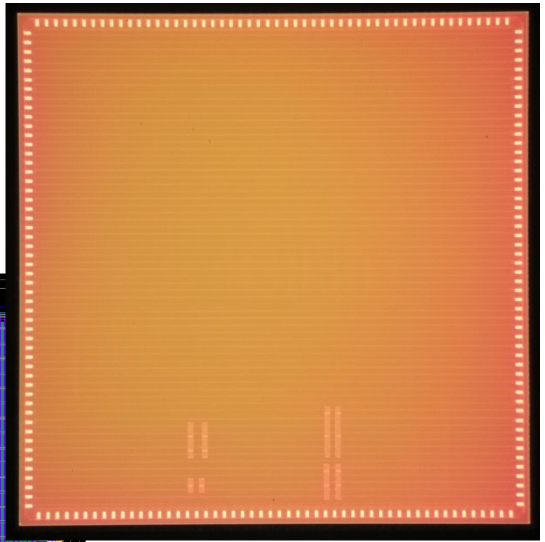
LSTM / MLP
5.2 x 4.2 mm

Sparse CNN
3.6 x 1.4 mm

Rocket Core
1.3 x 1.4 mm

Chipletized
Rocket Core
(with AIB)
3.2 x 1.6 mm





Fujitsu 55 nm

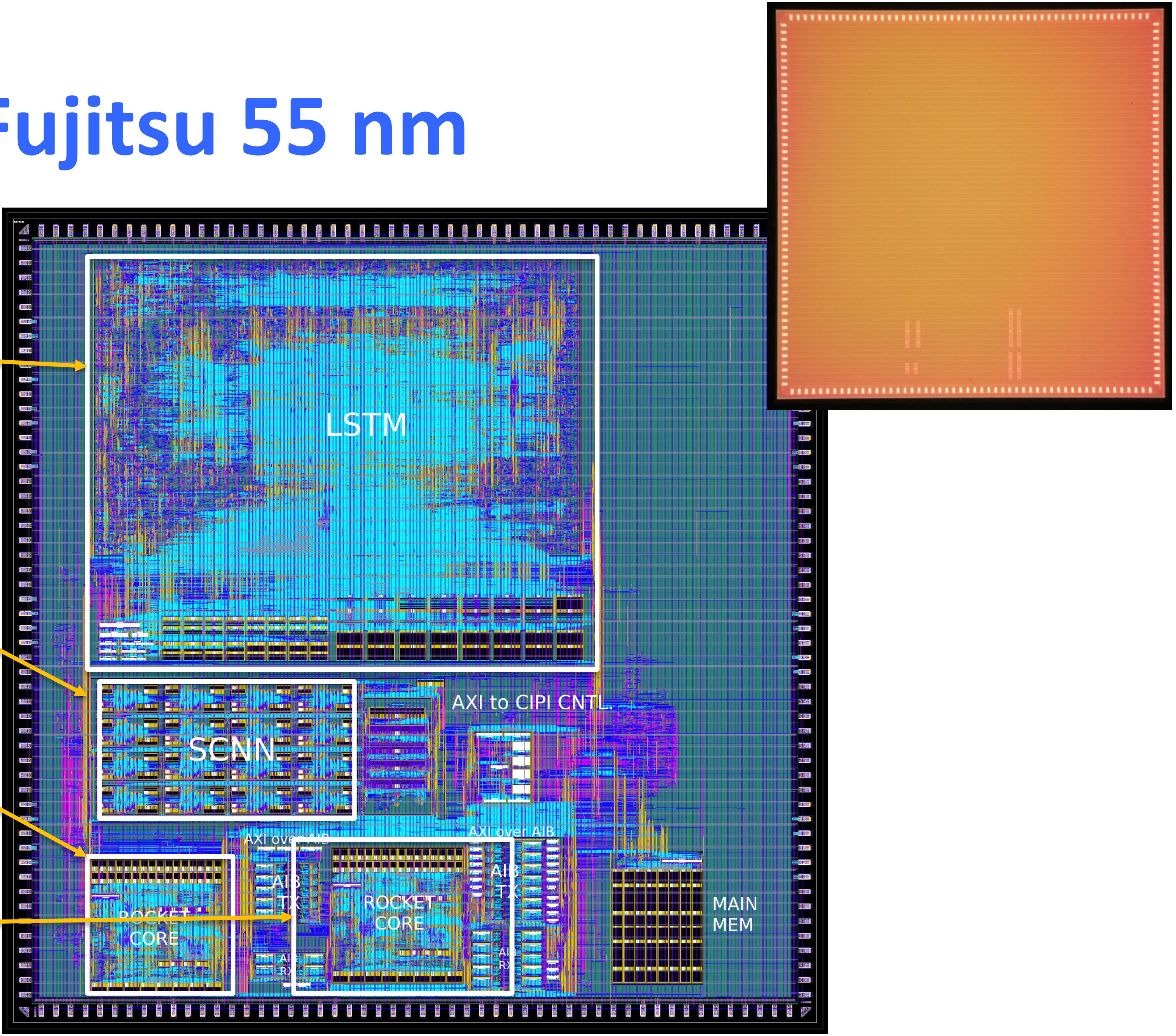
Simulates at 250 MHz

LSTM / MLP
5.2 x 4.2 mm

Sparse CNN
3.6 x 1.4 mm

Rocket Core
1.3 x 1.4 mm

Chipletized
Rocket Core
(with AIB)
3.2 x 1.6 mm


LSTM
SCNN
AXI to CIPI CNTL
AXI over AIB
AXI over AIB
AIB TX
AIB TX
ROCKET CORE
ROCKET CORE
AIB RX
AIB RX
MAIN MEM