National Science Foundation's Industry/University Cooperative Research (I/UCRC) Program

Embedded GPU Cluster Computing Framework for Inference of Convolutional Neural Networks

AFRL THE AIR FORCE RESEARCH LABORATORY
LEAD | DISCOVER | DEVELOP | DELIVER

Mission-Critical Computing
NSF CENTER FOR SPACE, HIGH-PERFORMANCE, AND RESILIENT COMPUTING (SHREC)

IEEE HPEC 2019 Presentation

University of Pittsburgh
BYU BRIGHAM YOUNG UNIVERSITY
Virginia Tech 1872
UF UNIVERSITY of FLORIDA

Dr. Andrew C. Pineda
U.S. Air Force Research Laboratory Space Vehicles Directorate

Evan Kain
University of Pittsburgh

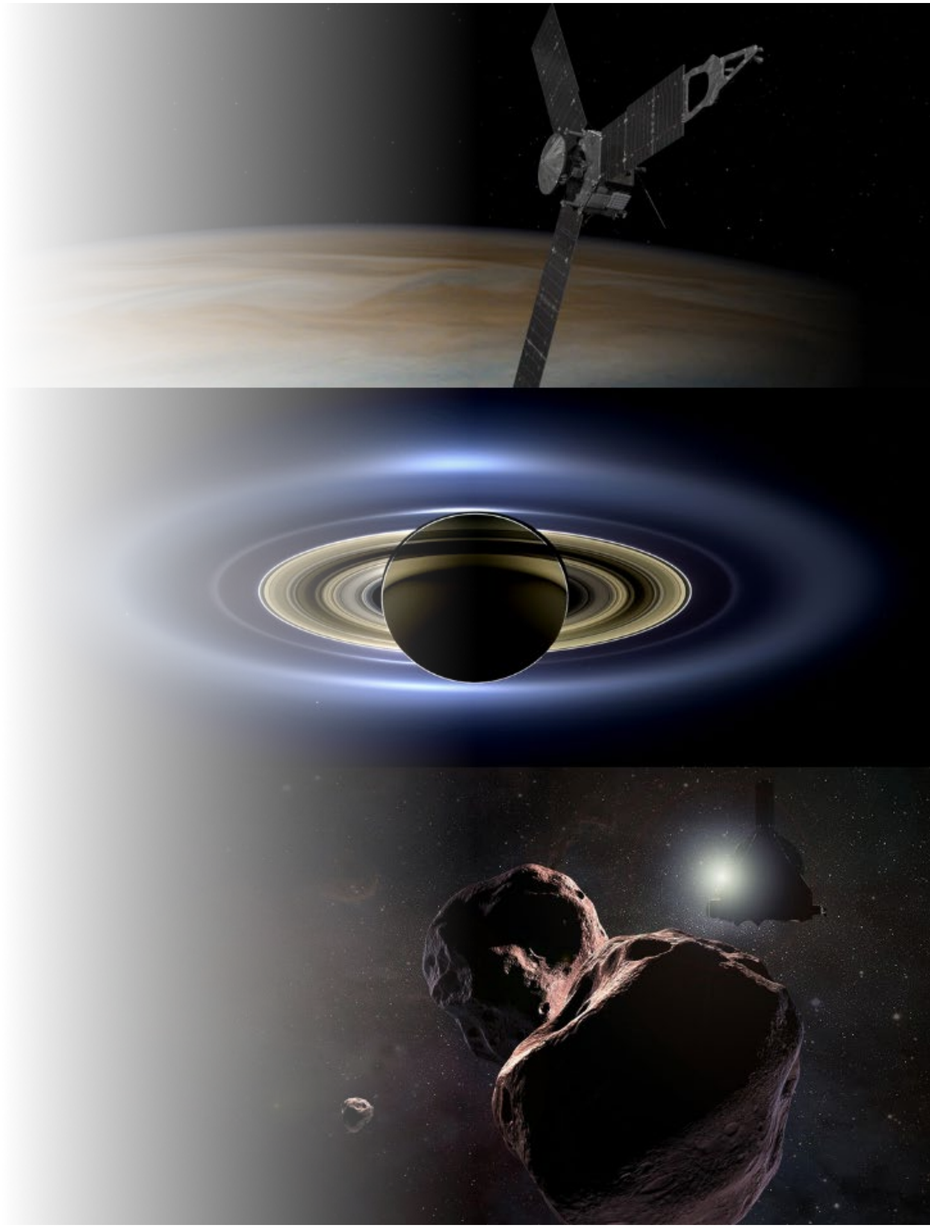
Diego Wildenstein
Arizona State University*

* Now attending Univ. of Pittsburgh

DISTRIBUTION A – Approved for public release: distribution unlimited. AFMC-2019-034, OPS-19-34980

Outline

- Goals, Motivations, and Challenges
- Background
- Approach
- Results
- Conclusions



Goals, Motivations, and Challenges

- **Goals**
  - Characterize speedup, parallel efficiency, and other scaling properties of MPI wrapper for embedded GPU SOCs
  - Predict best-case performance of TMR system for fault tolerance
- **Motivations**
  - Run complex apps without flying large, power-hungry GPUs
  - Match algorithms to space-compatible hardware
- **Challenges**
  - Non-trivial overhead for domain decomposition
  - Tradeoffs between efficiency and communications overhead

MPI: Message-Passing Interface
TMR: Triple Modular Redundancy

GPU: Graphics Processing Unit
SOC: System on a Chip

TX2 Board

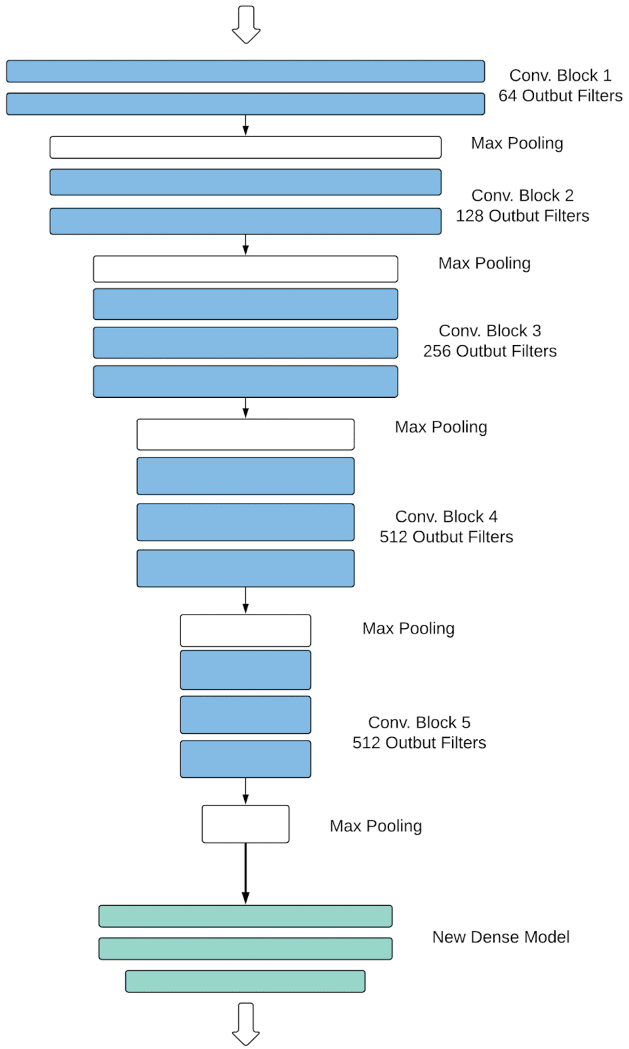
- **256 CUDA cores**
- **Quad-core ARM A57 CPU**
- **Power dissipation**
  - NVIDIA Tegra TX2 – 15W TDP
  - NVIDIA GeForce RTX 2080 Ti – 260W TDP
  - NVIDIA Tesla V100 – 250W TDP

Source: https://devblogs.nvidia.com/jetson-tx2-delivers-twice-intelligence-edge/

Source: https://www.nvidia.com/en-in/geforce/graphics-cards/rtx-2080-ti/

Convolutional Neural Network

- **Treated as black box in this study**
- **Sublinear increase in processing time as image size increases**
- **Space application**
  - Low power
  - Inherently reliable
  - On-orbit inference



Conv. Block 1
64 Outbut Filters

Max Pooling

Conv. Block 2
128 Outbut Filters

Max Pooling

Conv. Block 3
256 Outbut Filters

Max Pooling

Conv. Block 4
512 Outbut Filters

Max Pooling

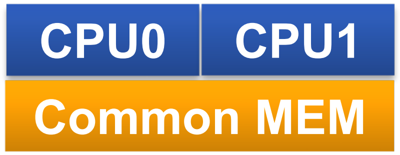
Conv. Block 5
512 Outbut Filters

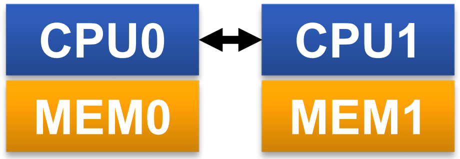
Max Pooling

New Dense Model

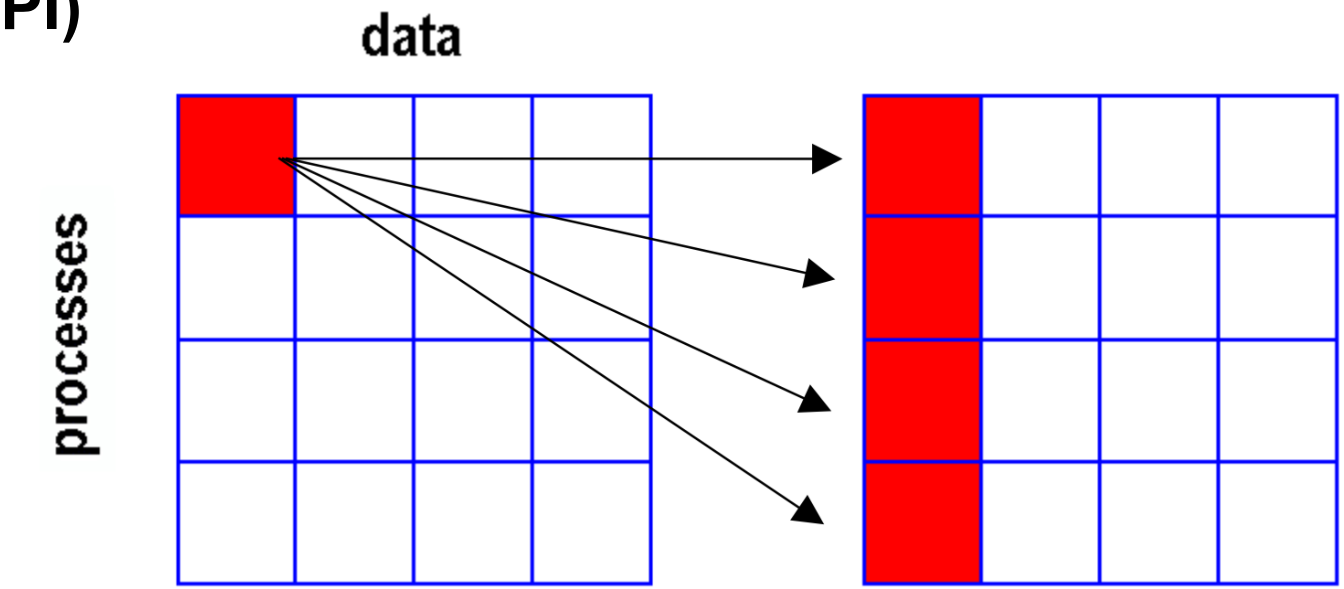
Source: https://www.codesofinterest.com/p/build-deeper.html

Distributed-Memory Multiprocessing



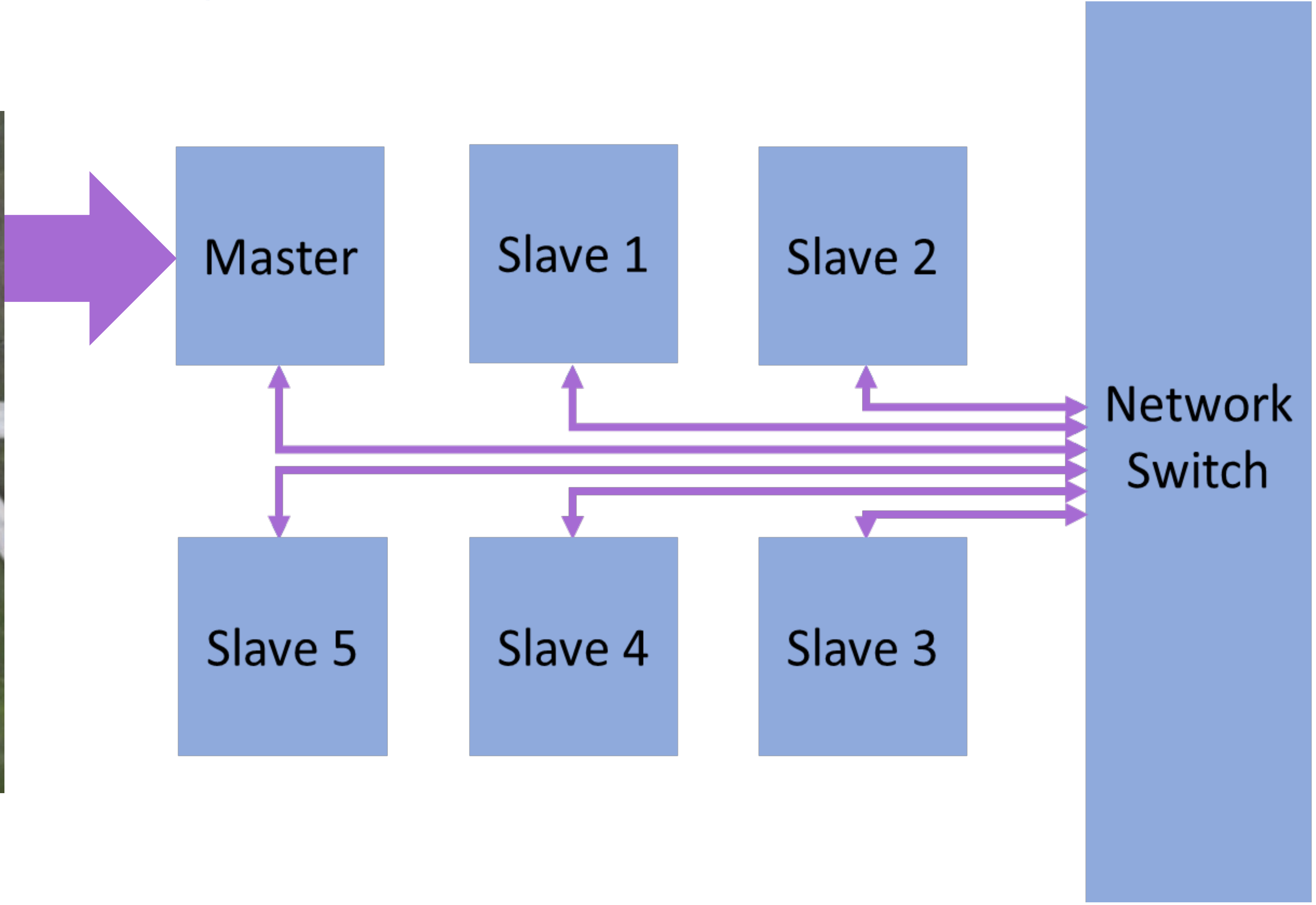


- **Distributed-Memory Model**
  - Multiple compute nodes carry out work
  - Nodes communicate and synchronize via function calls
  - Each node can contain multiple cores with internal shared memory
  - Contrast to shared memory: many threads on one device
- **Message-Passive Interface (MPI)**
  - Many implementations
    - MPICH
    - OpenMPI – de facto standard
    - Vendor specific implementations
  - Processes operate independently
  - Data passed in messages
  - Synchronization necessary



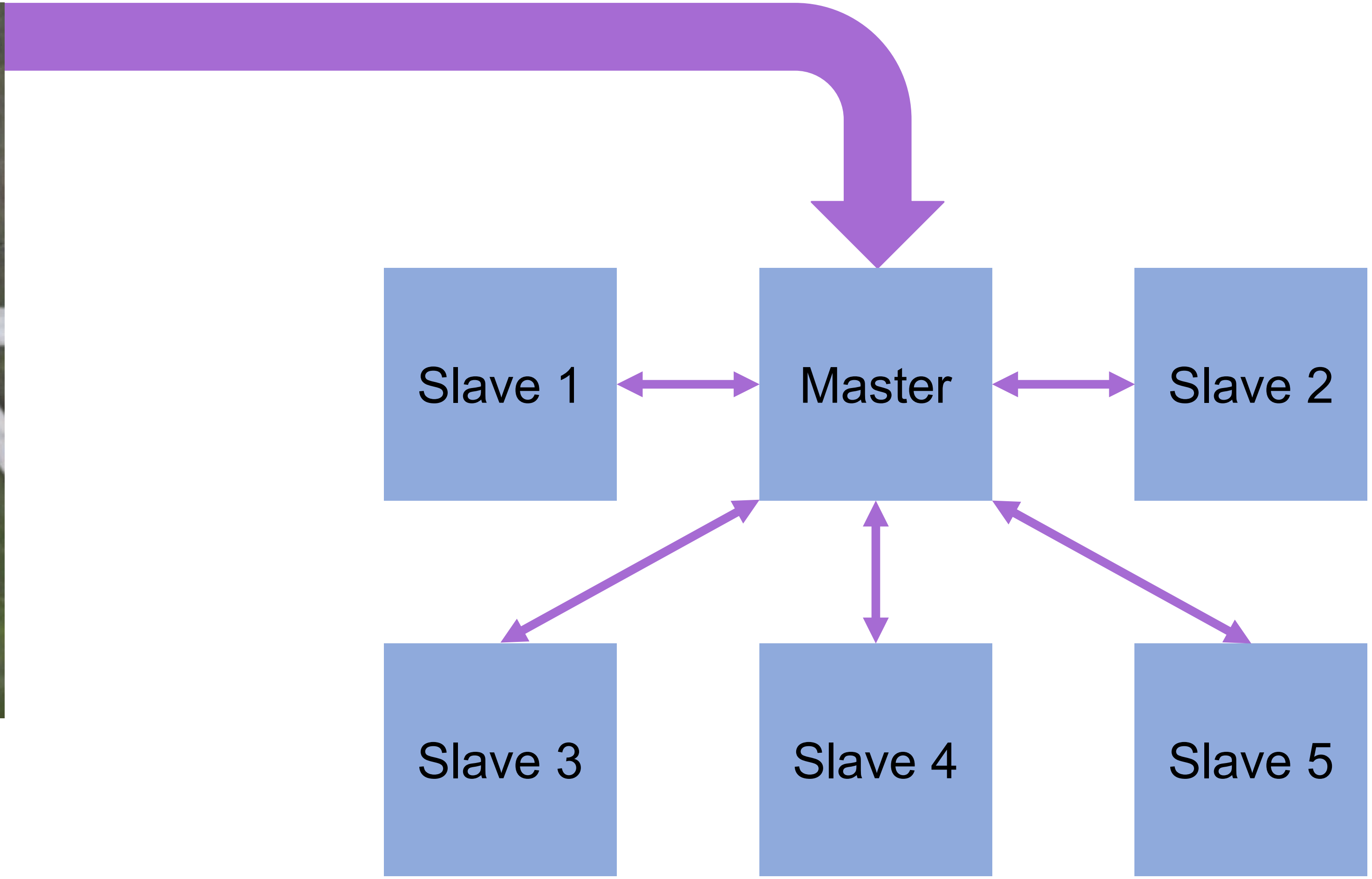
Physical Network Setup

Source: Google maps satellite view



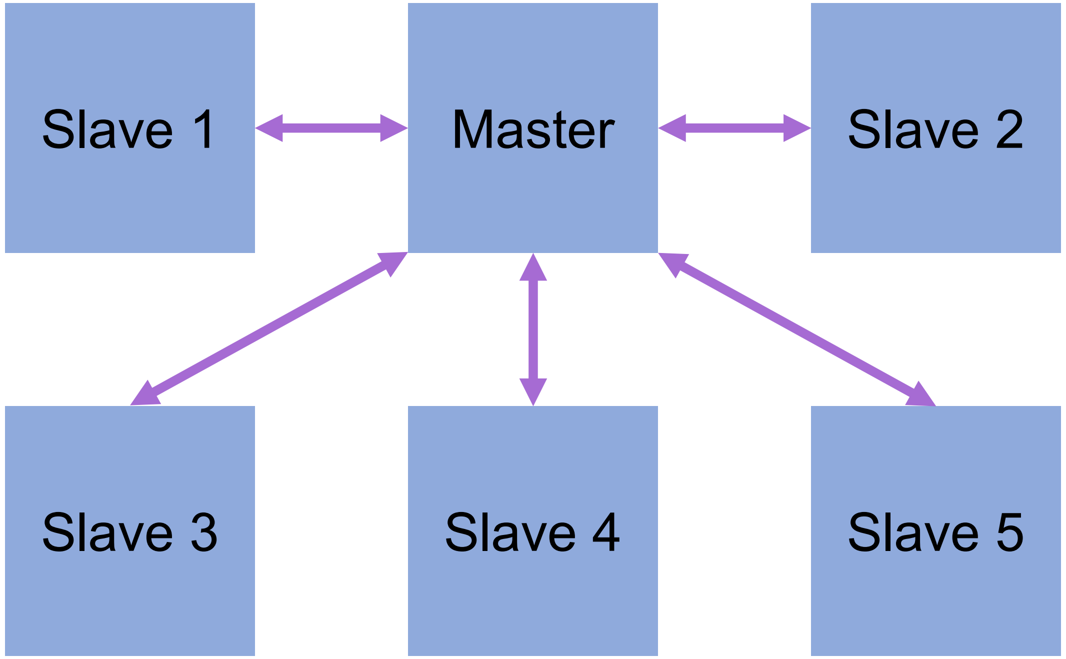
Logical Network Setup

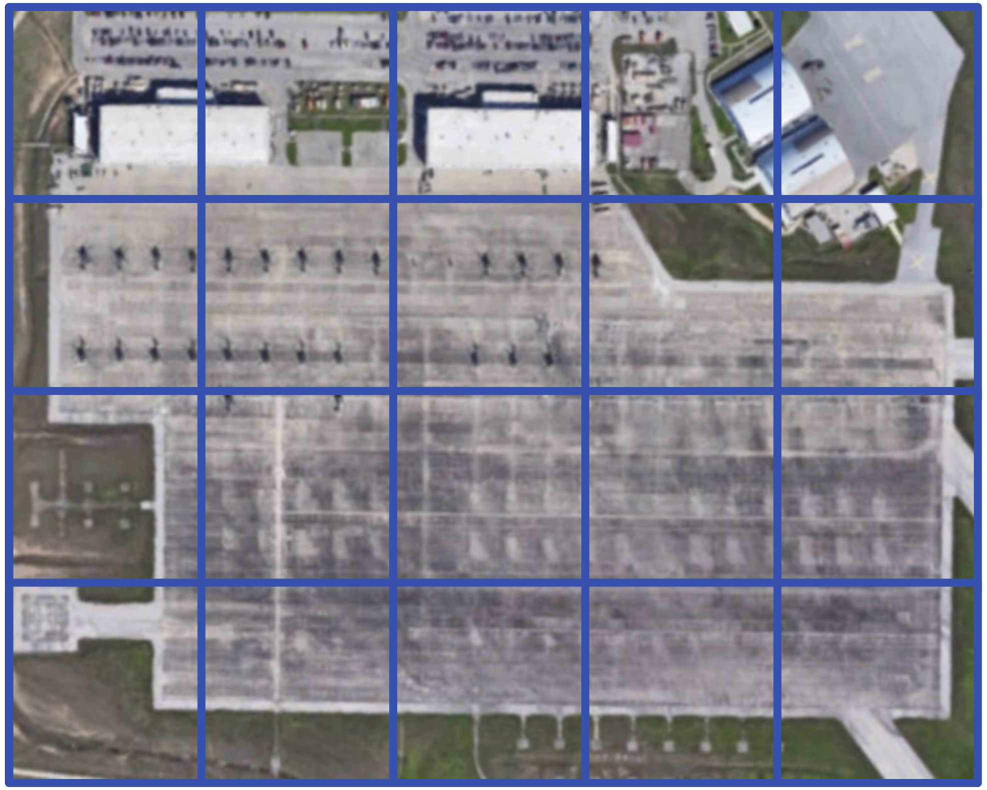
Source: Google maps satellite view



Application

- **Tiles may overlap to avoid splitting objects**

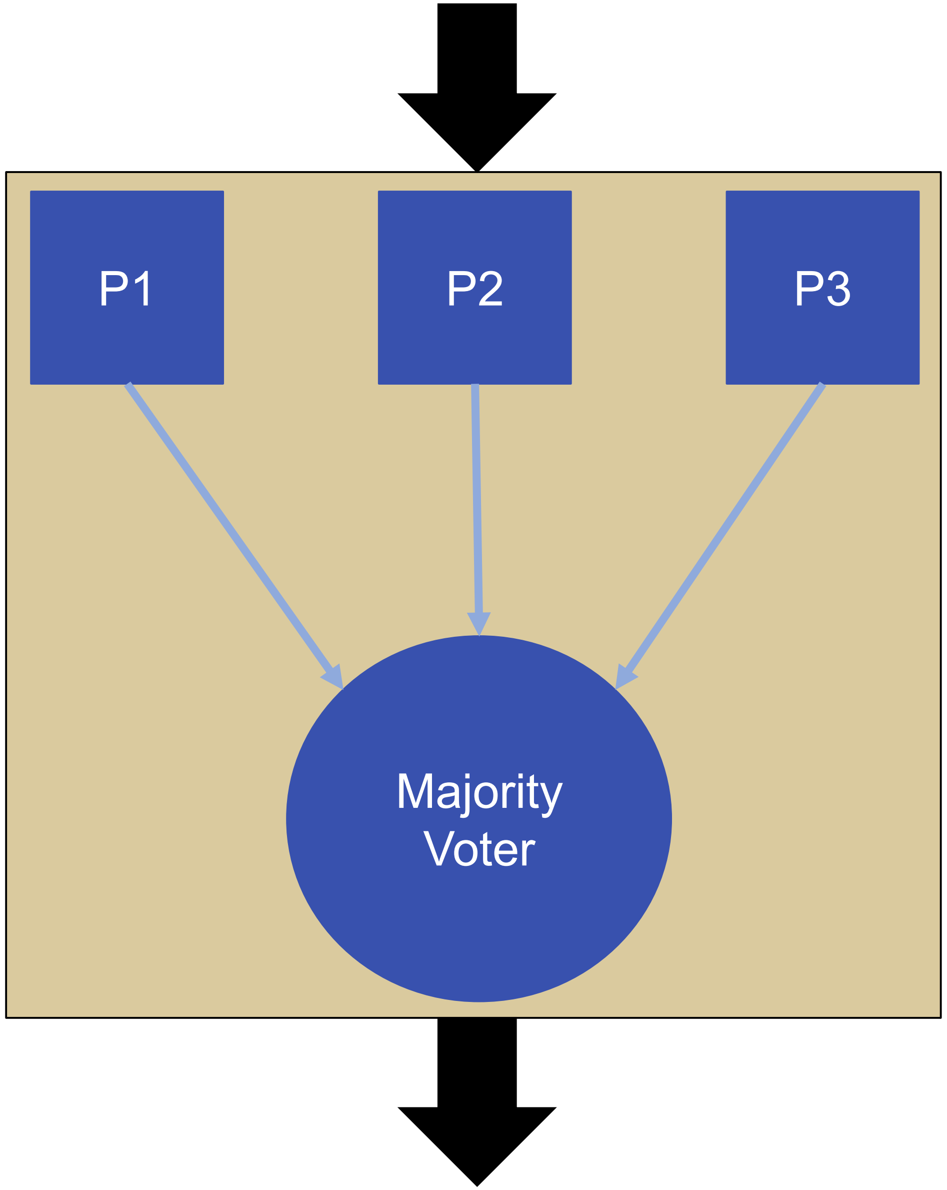




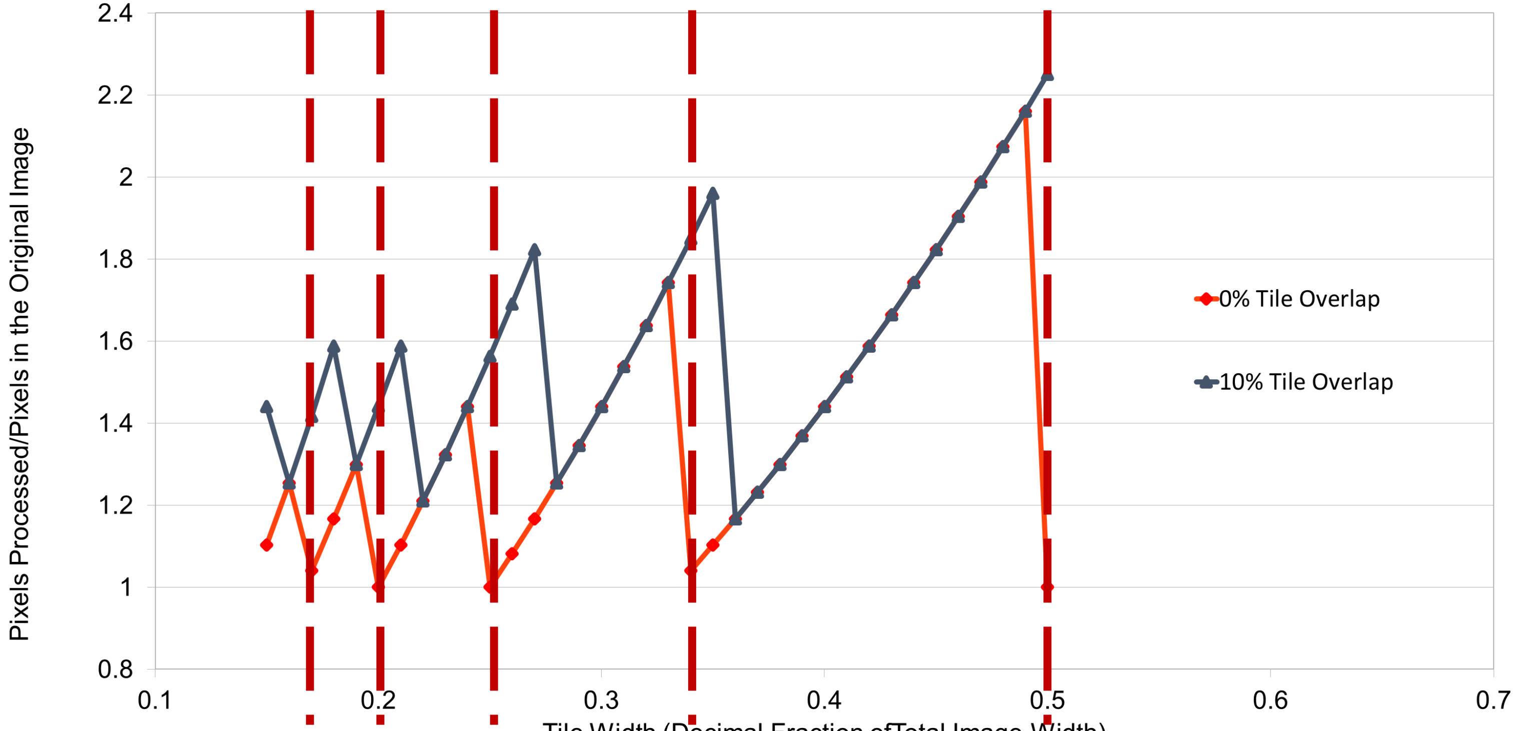
Source: Google maps satellite view

Evaluation Methodology

- **External timers**
- **Total program time**
  - Tiling
  - Inference
  - Translation of results
- **TMR performance model**
  - Not full TMR
  - Only model for most time-intensive parts
  - No use of fault-tolerant MPI

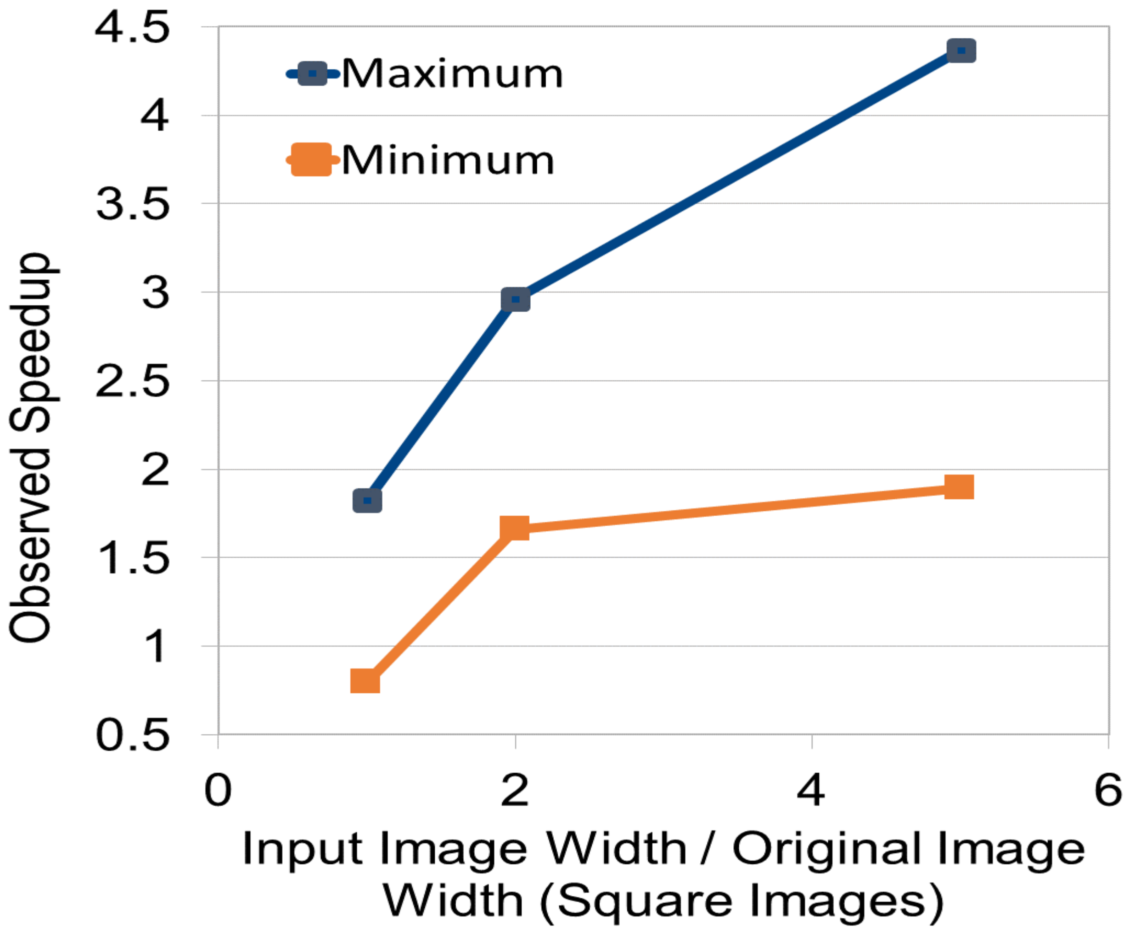


Performance of Tiling Scheme



- **Zero padding adds data**
  - Minima in additional data
  - Maxima in speedup
- **Tile width**
  - Optimal tile size
  - Load balancing better for small tiles
- **Overlap**
  - Incurs performance penalty
  - 10% case representative of large range of overlaps

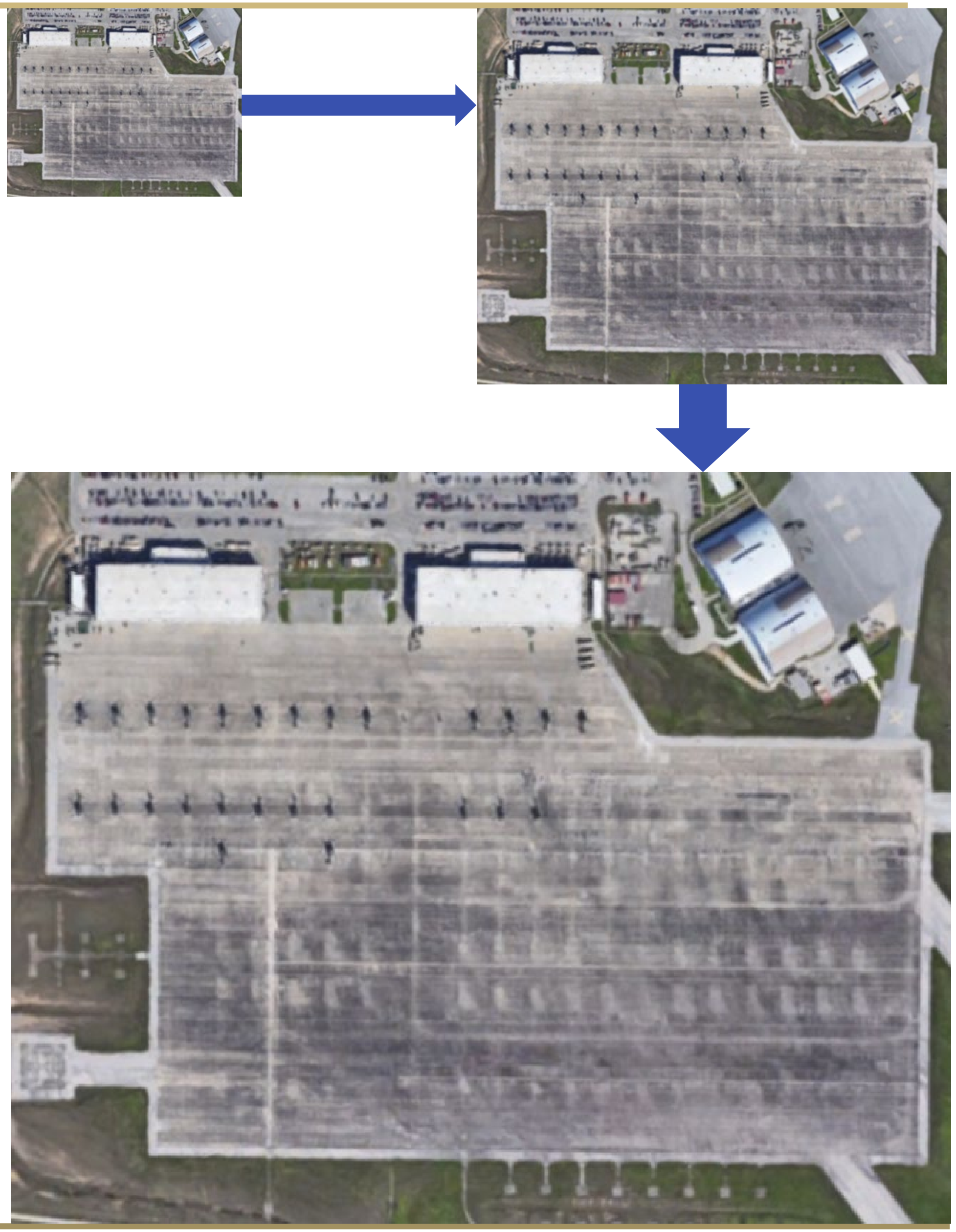
Parallel Performance



- **Linear decay in parallel efficiency**
- **Serial bottleneck**
- **Curve offset from overlap**
  - Overlap always generates extra data
  - Zero padding amplifies effect
  - No overlap is limiting case
- **Input image size**
  - Maximum speedup
  - Minimum speedup

Conclusions

- Maximum speedup parameters
  - 4.3 × speedup
  - 6 nodes
  - 0% tile overlap
  - $\frac{\mathit{Tile\ Width}}{\mathit{Input\ Image\ Width}} = 17\%$
- Maximum input image size increased
  - Successfully increased total area by factor of 244
  - Limited by routine used to read image
  - No theoretical limit to input image size
- Best-case TMR performance
  - 3× slower
  - No loss of accuracy for naïve fault model

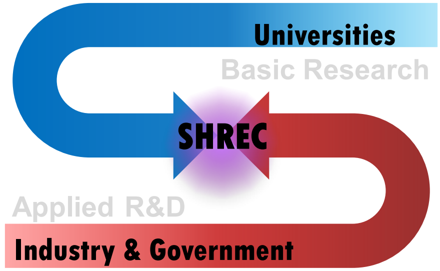


Acknowledgements

- **Mr. Joshua Donckels**
  - AFRL Space Vehicles Directorate
  - Assistance with CNN
  - Code review
- **AFRL Spacecraft Processing Architectures and Computing Environment Research (SPACER) Laboratory**
- **Universities Space Research Association (USRA)**
- **Dr. Alan D. George**
  - NSF SHREC
  - Advisement and review

NSF SHREC Center

- **NSF Center for Space, High-performance and Resilient Computing**
  - Founded in 2017 with focus on **mission-critical computing** needs
  - Intersection of **space, embedded, and high-performance** computing
  - **Four university sites** and over 30 industry and government partners



- **Formerly NSF Center for High-performance Reconfigurable Computing**
  - **CHREC** began operation in 2007 and was sunset in 2017
  - **NOVO-G** – large-scale reconfigurable supercomputer
  - **CHREC Space Processor (CSP)** – hybrid COTS/rad-hard CPU/FPGA space computer

- **For more information, please visit www.nsf-shrec.org**

This research was funded by industry and government members of the NSF SHREC Center and the National Science Foundation (NSF) through its IUCRC Program under grant number CNS-1738783.

QUESTIONS?