Exceptional service in the national interest

Sandia National Laboratories

CrossSim: GPU-Accelerated Simulation of Analog Neural Networks

T. Patrick Xiao, Christopher H. Bennett, Ben Feinberg, Matthew Marinella, Sapan Agarwal

Sandia National Laboratories, Albuquerque, NM
txiao@sandia.gov

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

LDRD LABORATORY DIRECTED RESEARCH & DEVELOPMENT

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

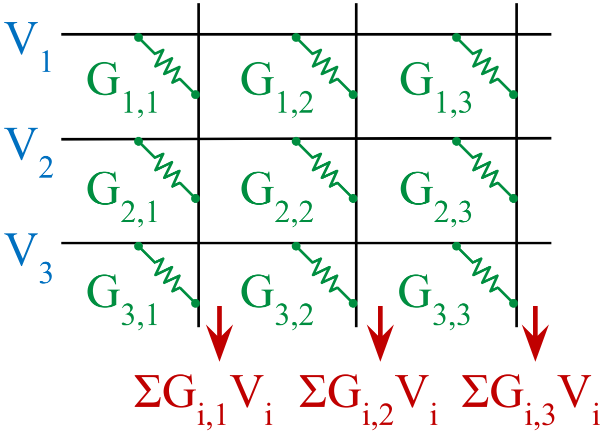
Deep learning inside memory arrays

Mathematical **Electrical**

Matrix-vector multiplication:

$\mathbf{Ax}$

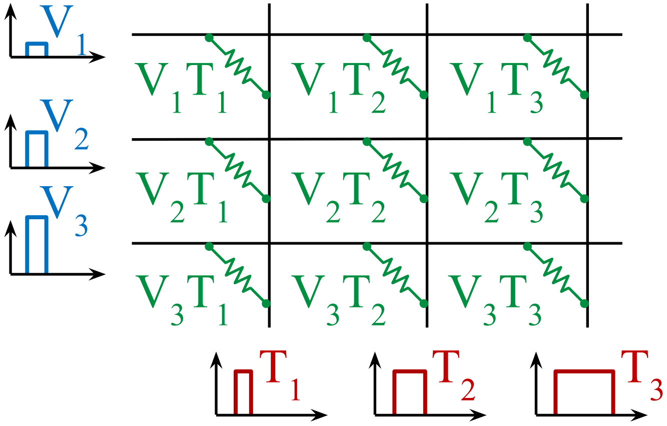
$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}^T \begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,1} & A_{3,2} & A_{3,3} \end{bmatrix} = \begin{bmatrix} \Sigma A_{i,1}x_i & \Sigma A_{i,2}x_i & \Sigma A_{i,3}x_i \end{bmatrix}$$



Outer product update:

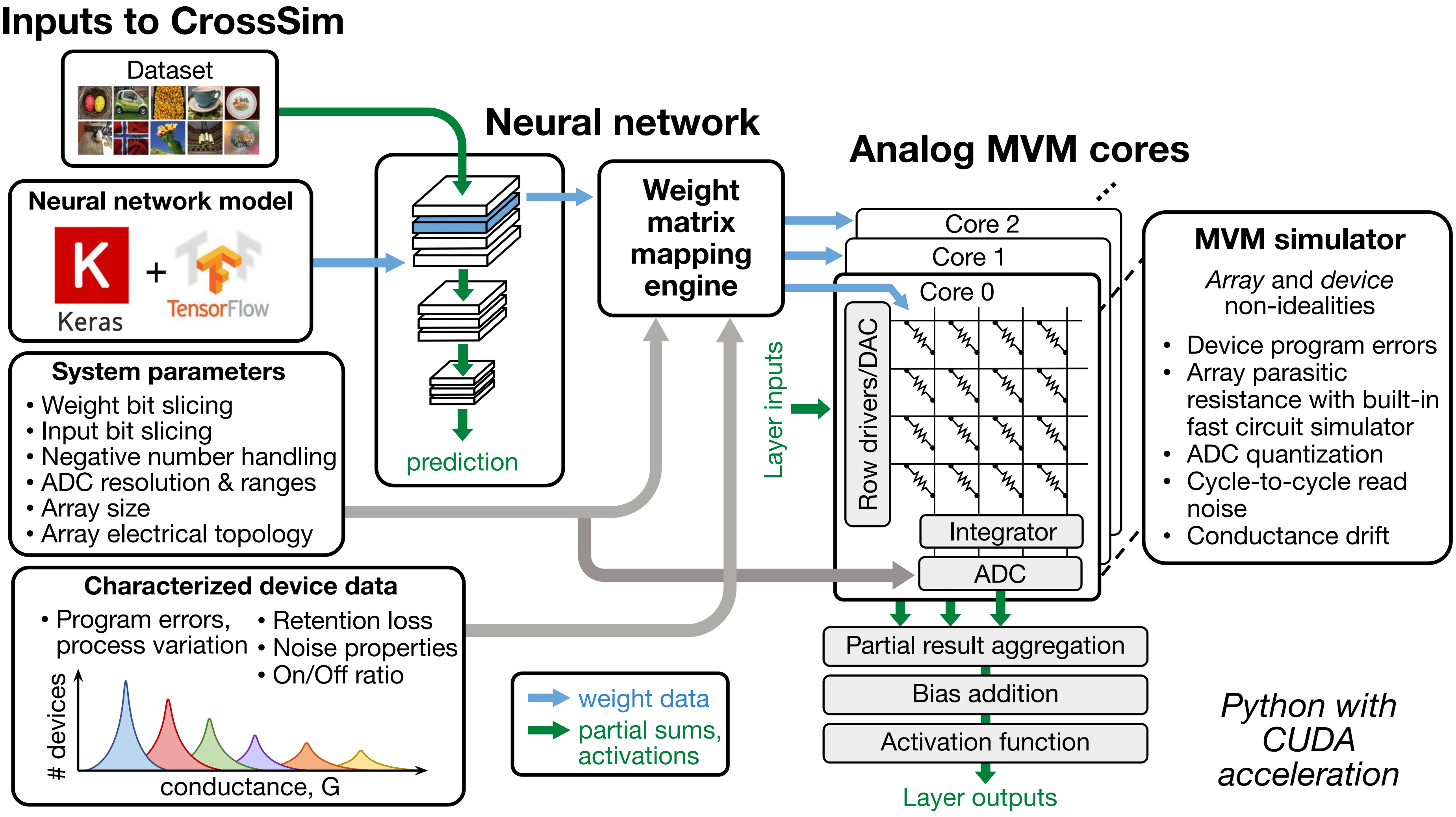
$\mathbf{x\delta^T}$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \otimes \begin{bmatrix} \delta_1 & \delta_2 & \delta_3 \end{bmatrix} = \begin{bmatrix} x_1\delta_1 & x_1\delta_2 & x_1\delta_3 \\ x_2\delta_1 & x_2\delta_2 & x_2\delta_3 \\ x_3\delta_1 & x_3\delta_2 & x_3\delta_3 \end{bmatrix}$$



Highly energy-efficient, *but is it accurate enough?*

CROSS SIM Inference



Inputs to CrossSim

Dataset

Neural network model

Keras + TensorFlow

System parameters

- Weight bit slicing
- Input bit slicing
- Negative number handling
- ADC resolution & ranges
- Array size
- Array electrical topology

Characterized device data

- Program errors, process variation
- Retention loss
- Noise properties
- On/Off ratio

Neural network

prediction

Weight matrix mapping engine

Layer inputs

Analog MVM cores

Core 2

Core 1

Core 0

Row drivers/DAC

Integrator

ADC

Partial result aggregation

Bias addition

Activation function

Layer outputs

MVM simulator

Array and *device* non-idealities

- Device program errors
- Array parasitic resistance with built-in fast circuit simulator
- ADC quantization
- Cycle-to-cycle read noise
- Conductance drift

weight data

partial sums, activations

Python with CUDA acceleration

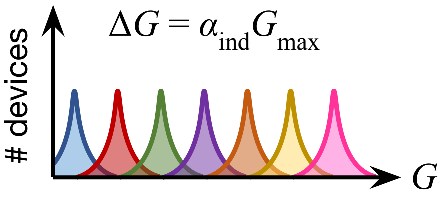
To be released soon! Check cross-sim.sandia.gov

Multi-scale modeling of inference accuracy

Device properties affect accuracy

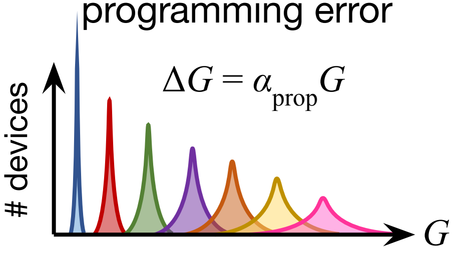
State-independent programming error

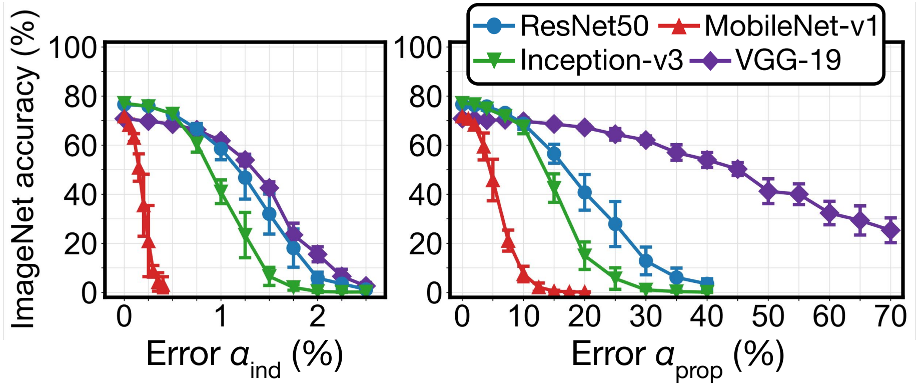
$\Delta G = \alpha_{ind} G_{max}$



State-proportional programming error

$\Delta G = \alpha_{prop} G$



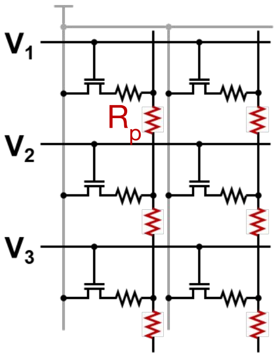


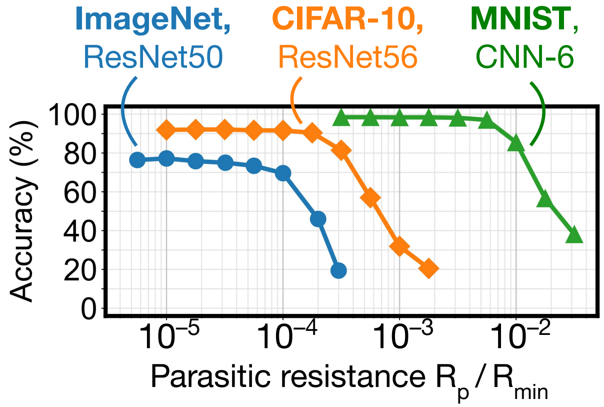
Xiao et al, *arXiv*:2109.01262, 2021

Xiao et al, *Semi Sci Tech*, Accepted (in press), 2021

Array design affects accuracy

CrossSim's fast built-in circuit simulator





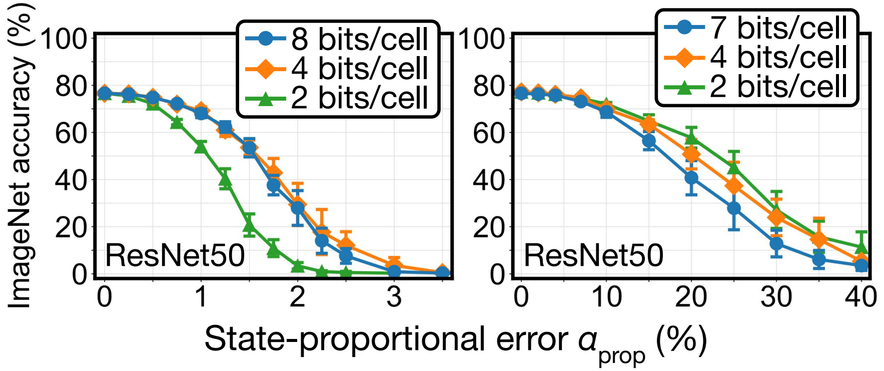
System architecture affects accuracy

Offset subtraction

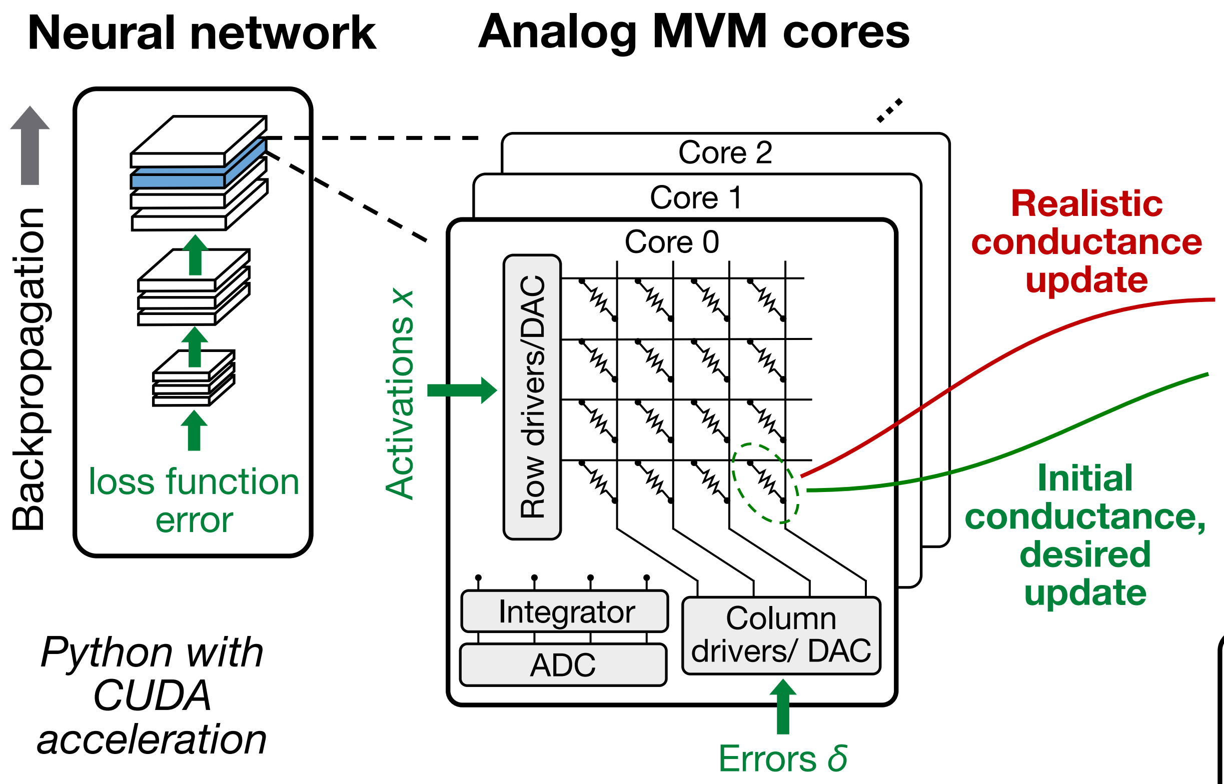
$W_{ij} \sim G_{ij} - G_{offset}$

Differential cells

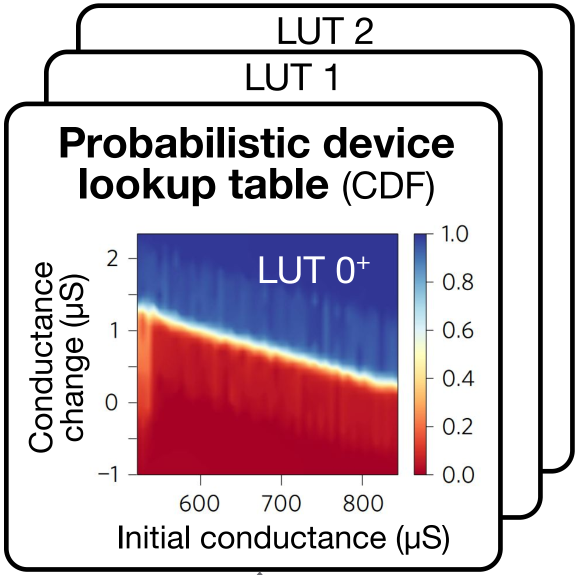
$W_{ij} \sim G^+_{ij} - G^-_{ij}$

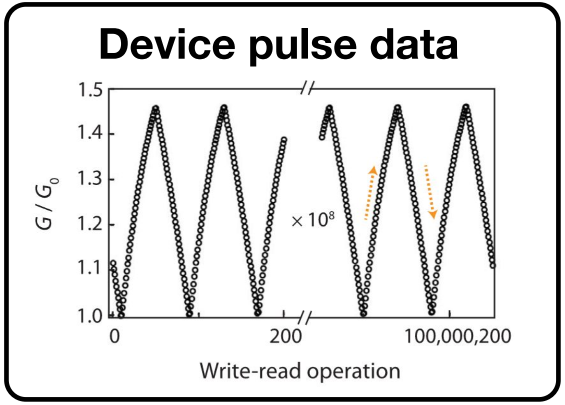


CROSS SIM Training



Python with CUDA acceleration





Lookup tables can model:

- Arbitrary device update **nonlinearity** and **asymmetry** properties, not describable by analytical equations
- Cycle-to-cycle **write noise**
- Device-to-device **variation**

Fuller et al, *Science* 2019
Bennett et al, *IRPS* 2019

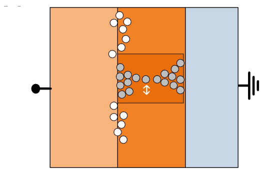
From device measurements to accuracy

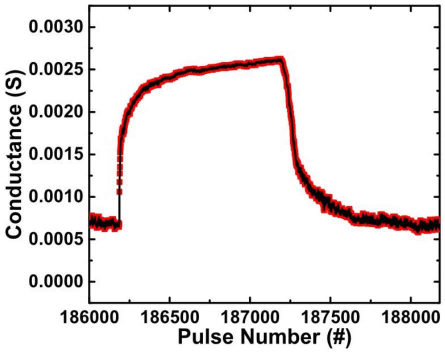
Device

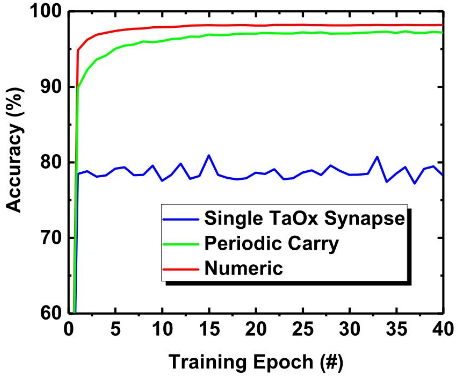
Pulse data

MNIST accuracy (2-layer MLP)

TaO_x ReRAM

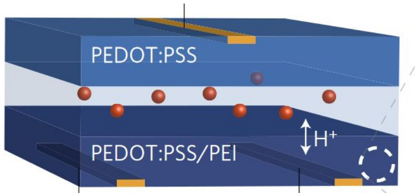


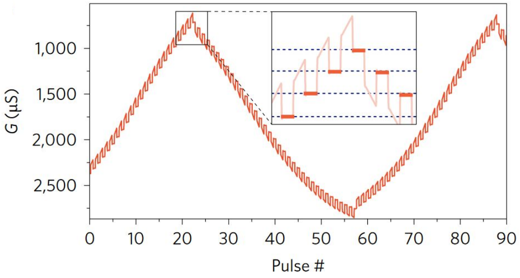


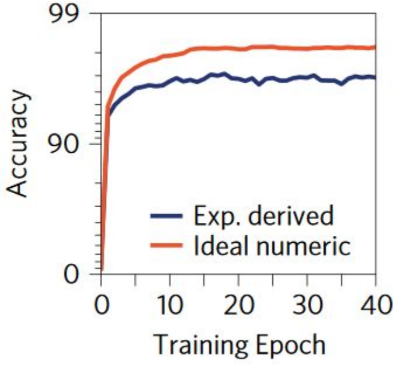


Marinella, Agarwal et al, *JETCAS* 2018

Electrochemical RAM

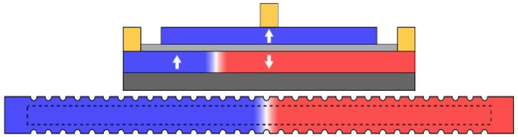


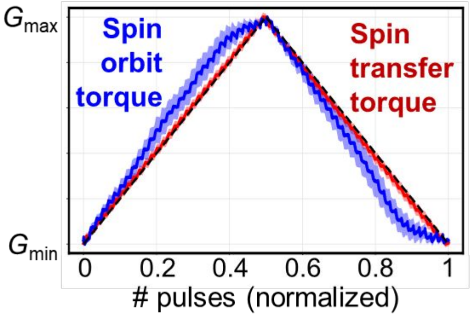


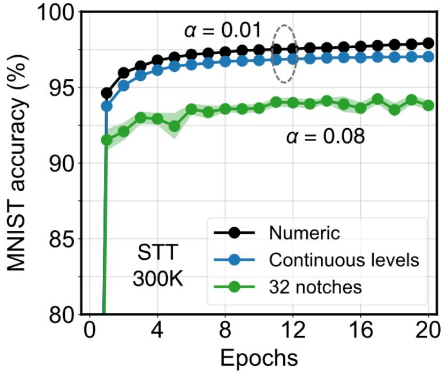


Van der Burgt et al, *Nature Materials* 2017

Domain wall magnetic tunnel junction







Liu et al, *Appl Phys Lett*, 2021