On Monitoring and Resiliency for Machine-Learning-Based Autonomous Systems

Michael Paulitsch

2021-06-25

IFIP WG 10.4 Talk

intel.

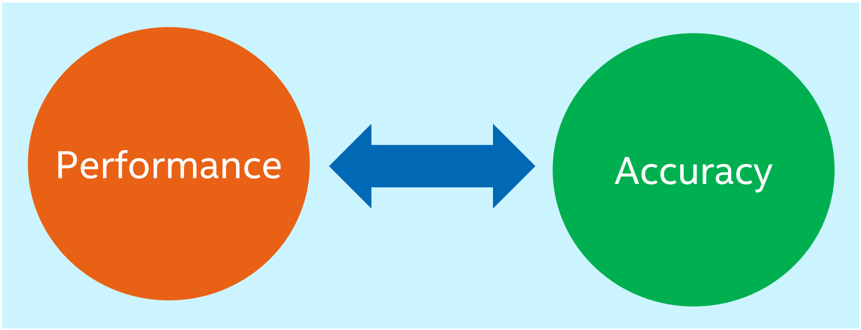
intel. labs

Overview of Two Related Topics

1. Resiliency: dependable AI/ML considering platform faults
2. Monitoring: safe perception

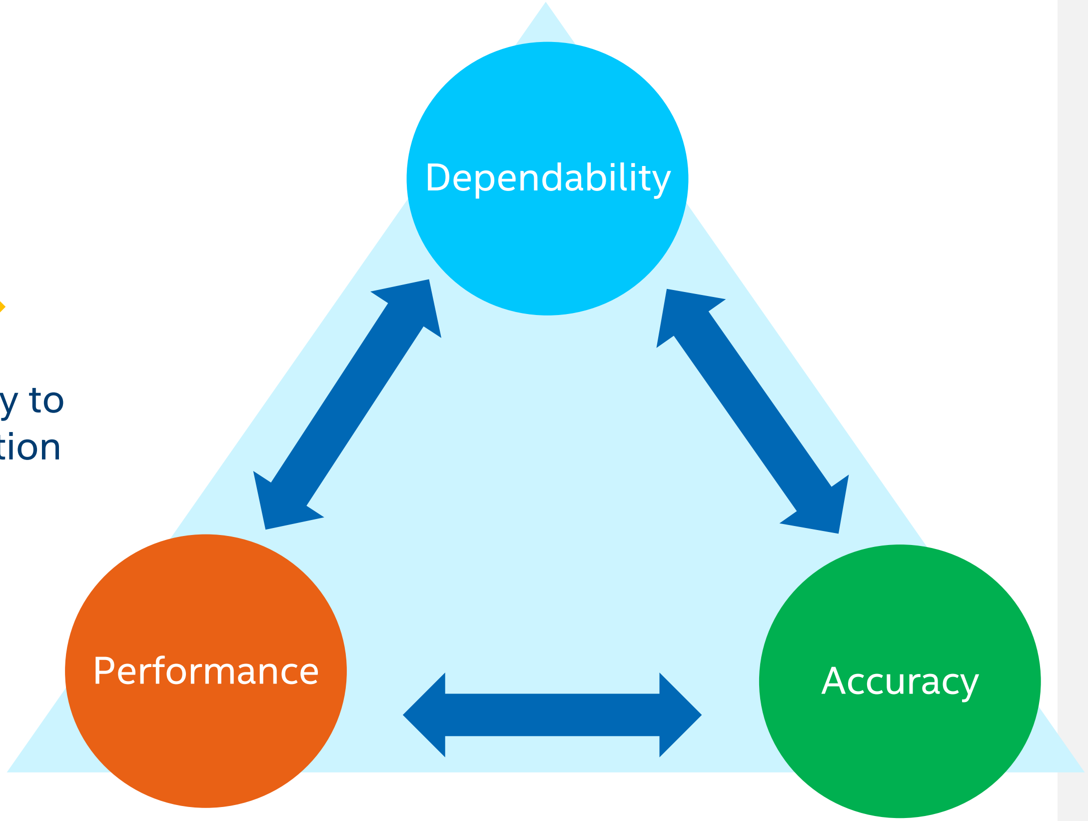
Accelerators

Dependability as Integral Part of Machine Learning System Optimization



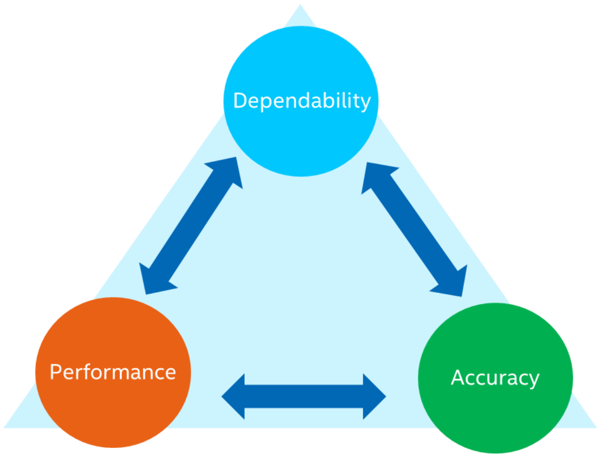
State-of-the-Art System Optimization with
Performace vs. Accuracy Trade-Off
(e.g., float32 => bfloat16)

Add Dependability to
System Optimization

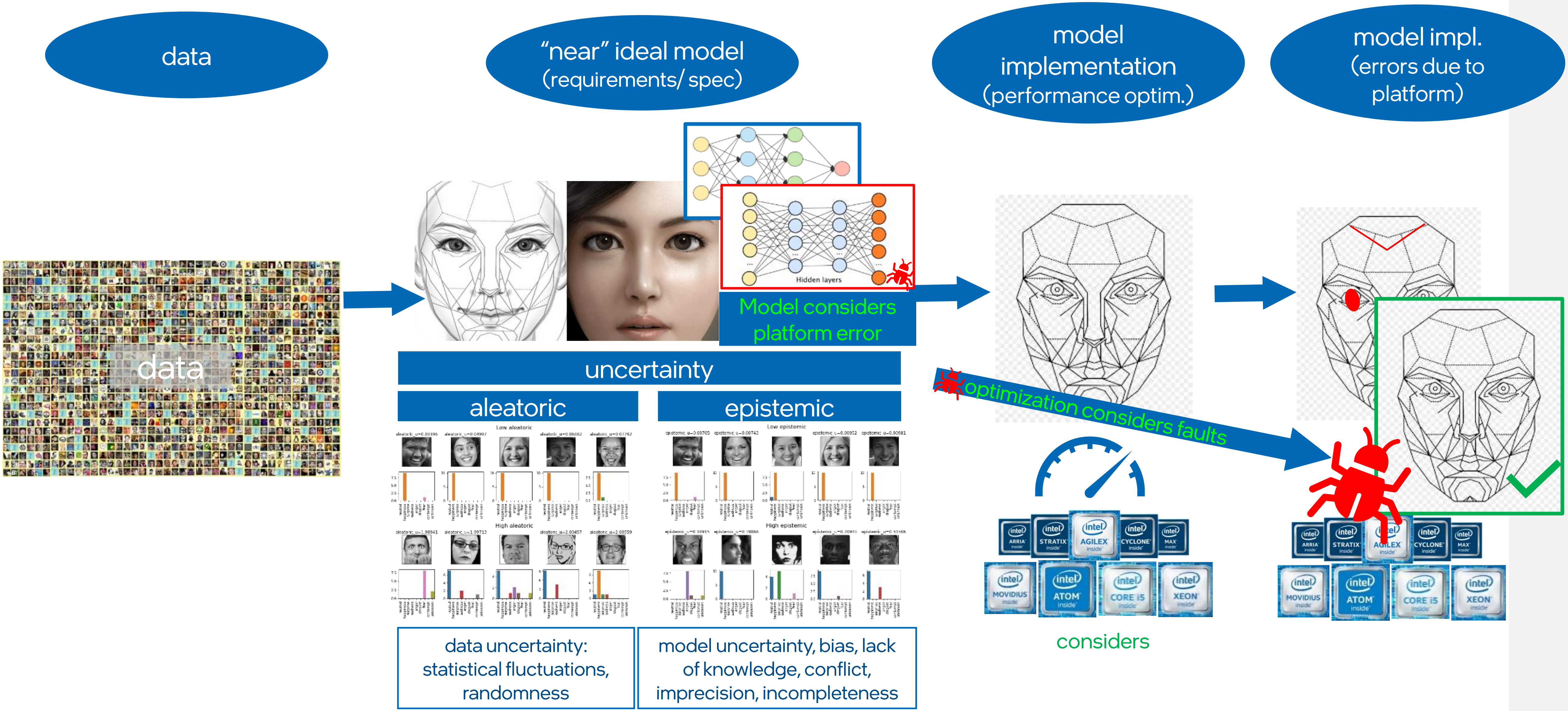


Research Questions?

- Dependability:
  - What are the faults and failure modes to consider?
  - Is dependability really a problem in machine learning?
  - What is the target? 10^{e-9} range for hazards? Failures due to silent data errors?
  - Can we trust AI/ML at all? Diversity arguments? Safety monitoring?
- Cost: overhead runtime and development
  - Can it be nearly free?
  - How to automate?
- Balance: the quest for win-win situation
  - Is there an optimal balance?
  - Monitoring versus generalizability?

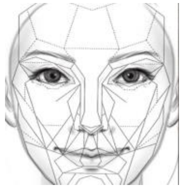


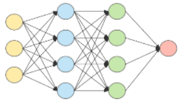
From Data to DNNs (CNNs) to Execution

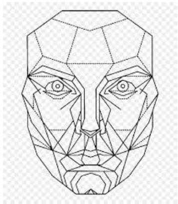


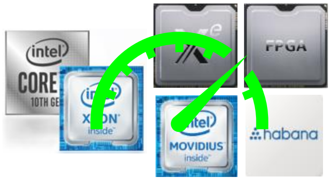
Summary

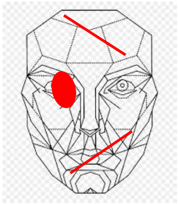
Data	Model Creation (requirements)	Model Optimization (performance)	Model Execution (platform errors)

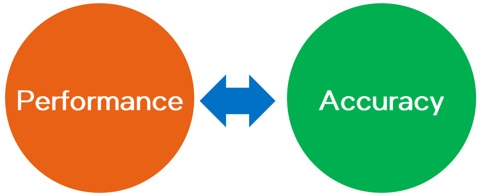






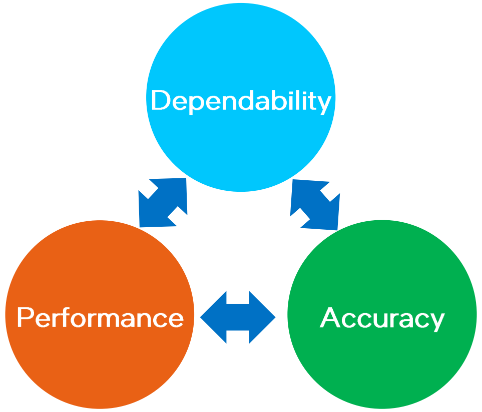






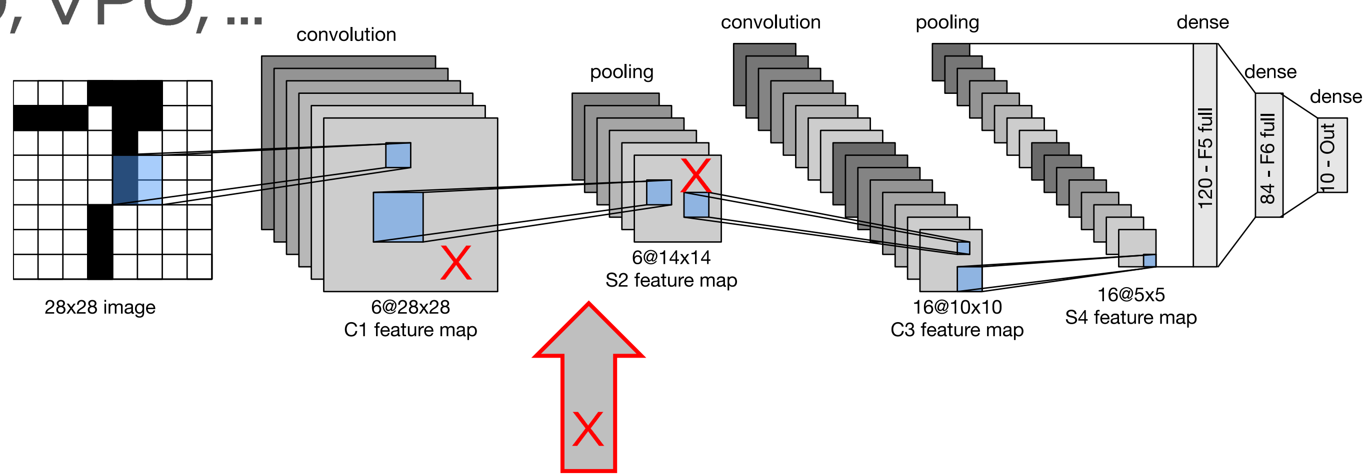
State-of-the-art system optimization with performance vs. accuracy trade-off (e.g., fp32 => bfloat16)

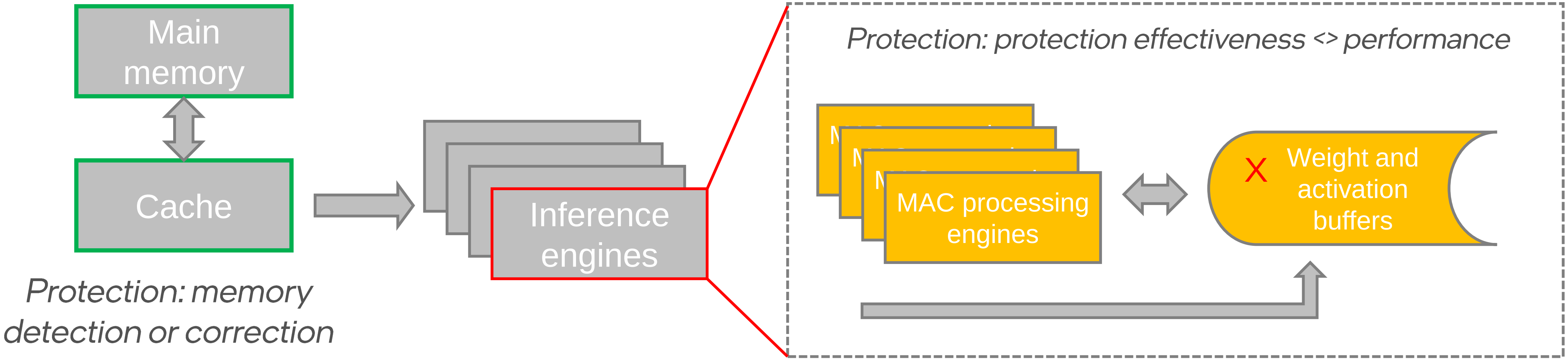
add dependability to system optimization



Common Elements in Intel CNN AI/ML accelerator HW

CPU extensions, GPGPU, VPU, …





Accelerators in Practice
Tiger Lake and Ice Lake

- Tiger Lake with Intel X^e accelerator
- Ice Lake: Advanced Vector Extensions (AVX)

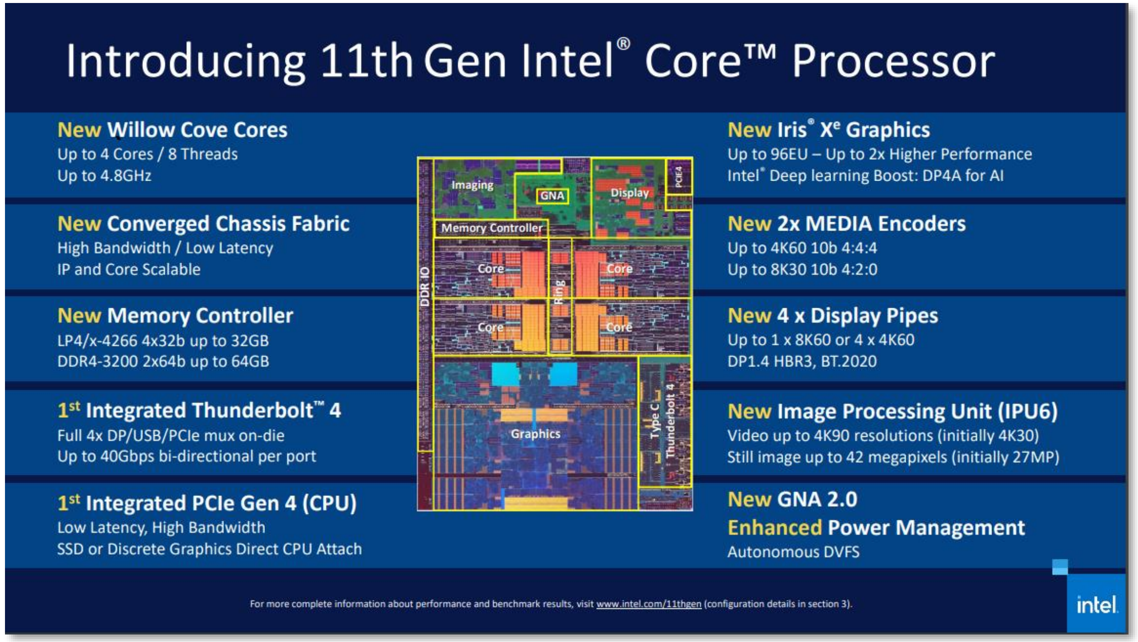
3rd Gen Intel® Xeon® Scalable Platform

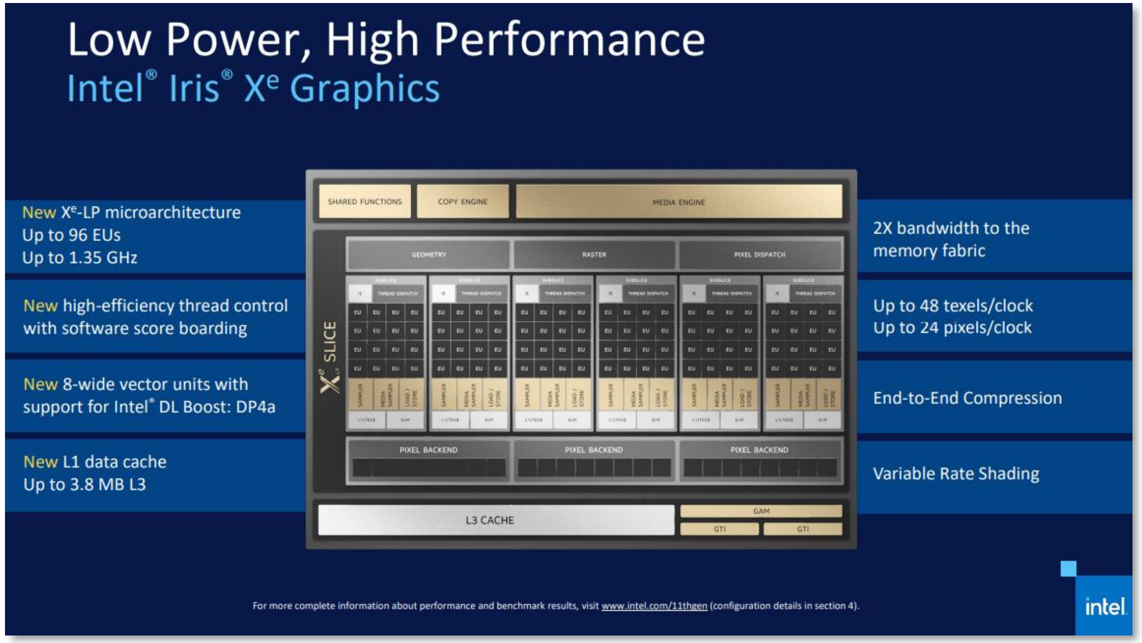
Feature	2nd Gen Intel® Xeon® Scalable Processor (Cascade Lake)	3rd Gen Intel® Xeon® Scalable Processors (Ice Lake)	Notes
Cores per Socket	4-28	8-40	New Sunny Cove architecture
L1/L2/L3 cache per core	32KB/1MB/1.375MB	48KB/1.25MB/1.5MB	Larger caches to enable fast access to data
Memory Channels and DIMM Speed	6 Up to 2933	8 Up to 3200	Huge boost in memory bandwidth & support for Intel® Optane™ PMem 200
Processor Interconnect: UPI links, speed	2 or 3, 10.4 GT/s	2 or 3, 11.2 GT/s	Improved bandwidth between processors
PCIe lanes per socket	PCIe 3.0, 48 Lanes (x16, x8, x4)	PCIe 4.0, 64 lanes (x16, x8,x4)	2x bandwidth and more PCIe lanes to support new Gen 4 SSD, Ethernet and other adjacencies
Workload Acceleration Instructions	AVX-512 VNNI DDIO	AVX-512, VNNI, DDIO vAES, vPCLMULQDQ, VPMADD52, VBMI, PFR, Crypto, SHA extensions, TME, SGX	Enable new capabilities and speedup performance
Platform Adjacencies		Intel® Optane™ PMem 200 series, Intel® Optane™ P5800X SSD, Intel DC P5510 SSD, Intel E810-C ethernet	

Designed to Move Faster, Store More, Process Everything

Performance made flexible. 37

source: 3rd Gen Intel Xeon Processor on intc.com

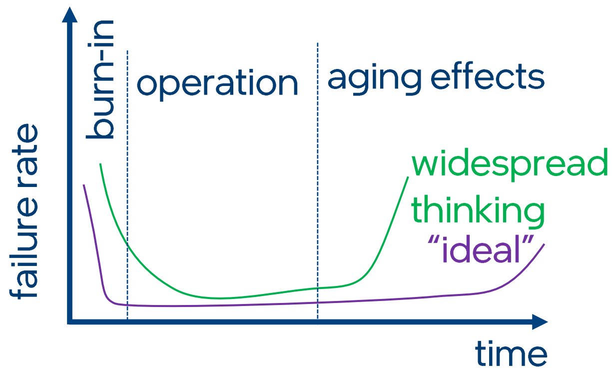




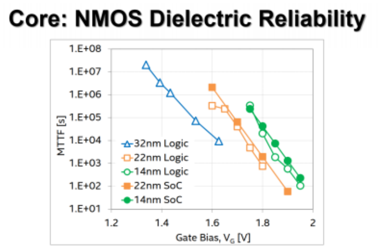
source: 11th Gen Intel Core Processor on intc.com

Fault / Error / Failures

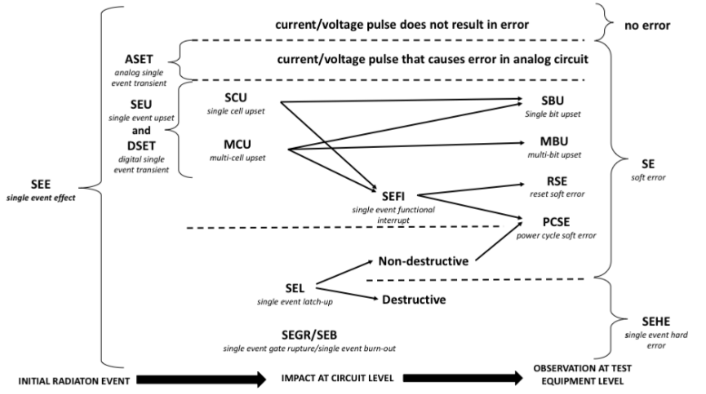
Hardware Faults – Error - Failures



Sightings / errata / known issues database



- SoC: same reliability as corresponding Logic node
- Strong reliability gains on both Tri-gate technologies



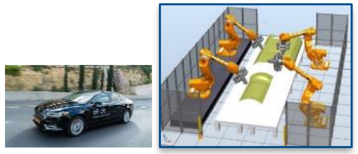
PERMANENT	INTERMITTENT	TRANSIENT

STUCK-AT ERROR	OPEN CIRCUIT ERROR	BRIDGING ERROR	SINGLE EVENT HARD ERROR (LATCHUP)	COMPLEX ERRORS	TRANSIENT OR HARD ERROR	SINGLE EVENT TRANSIENT, UPSET	MULTIPLE CELL UPSET	MULTIPLE BIT UPSET

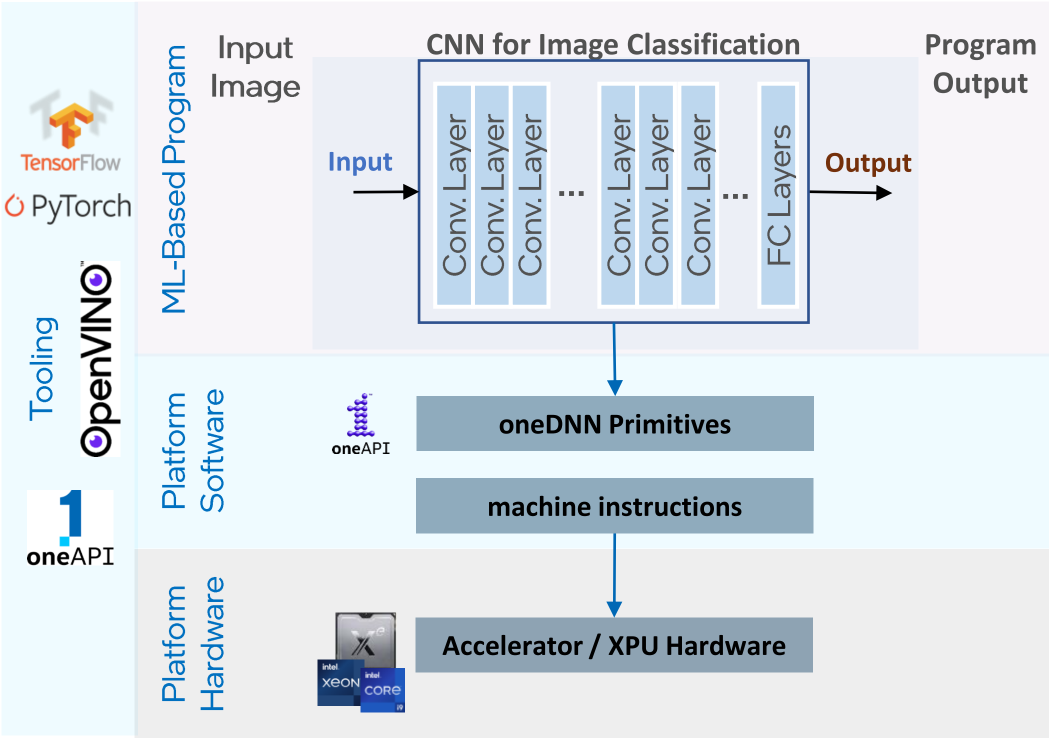
production (reliability, DPM)	design errors	aging	external - radiation

Dependable AI / ML - Resiliency

High-Performance Computing



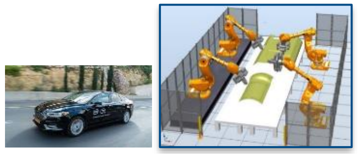
Safety-critical Applications



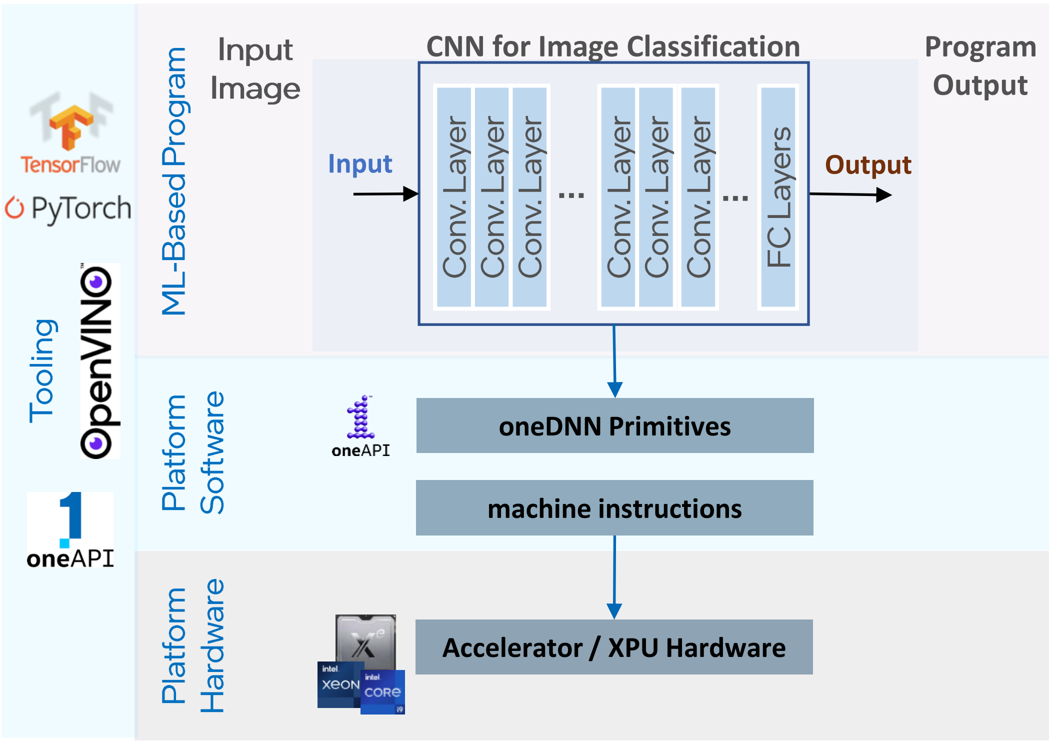
System View

Dependable AI / ML - Resiliency

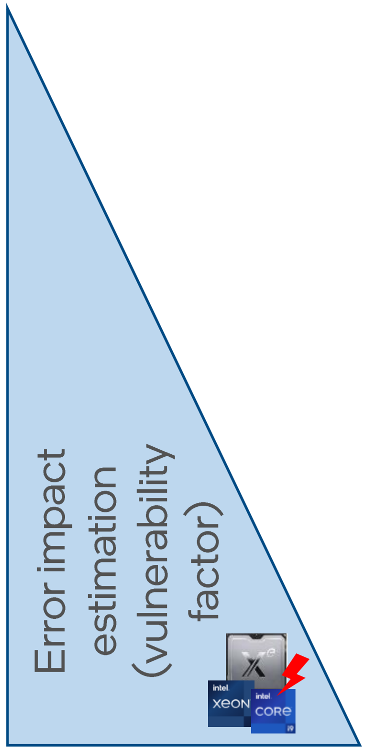
High-Performance Computing



Safety-critical Applications



System View



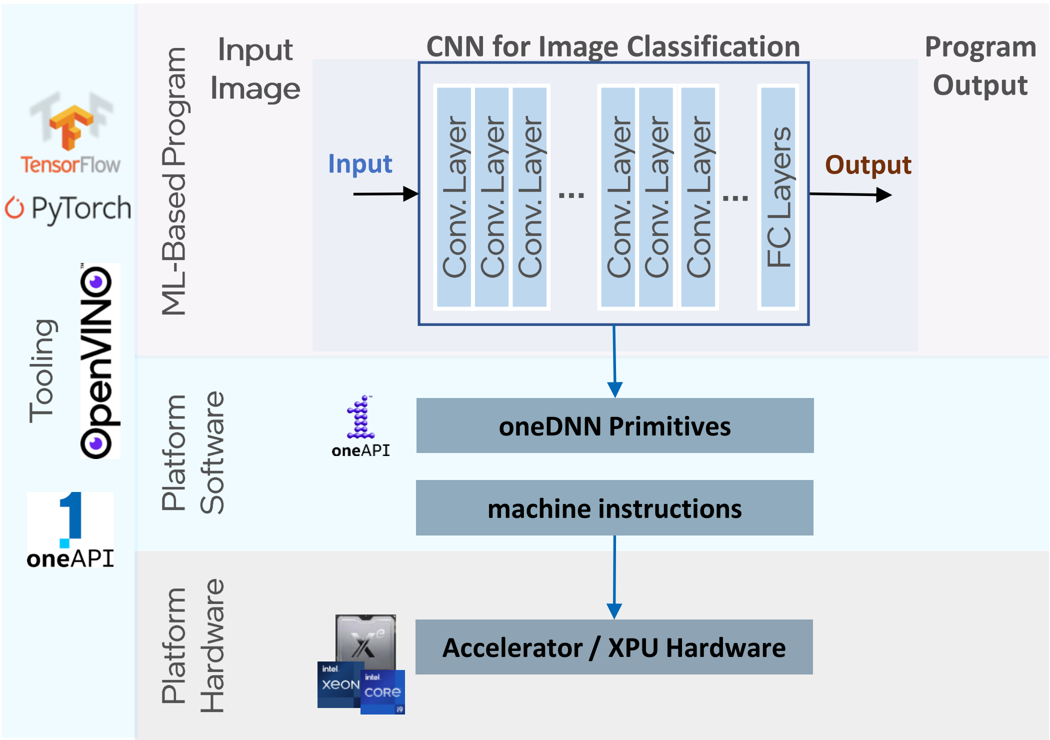
Dependability Threats

Dependable AI / ML - Resiliency

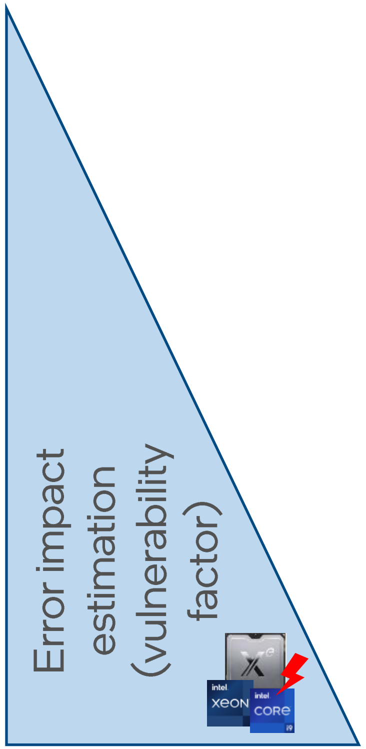


High-Performance Computing   Safety-critical Applications

dependability / reliability target



System View



Dependability Threats

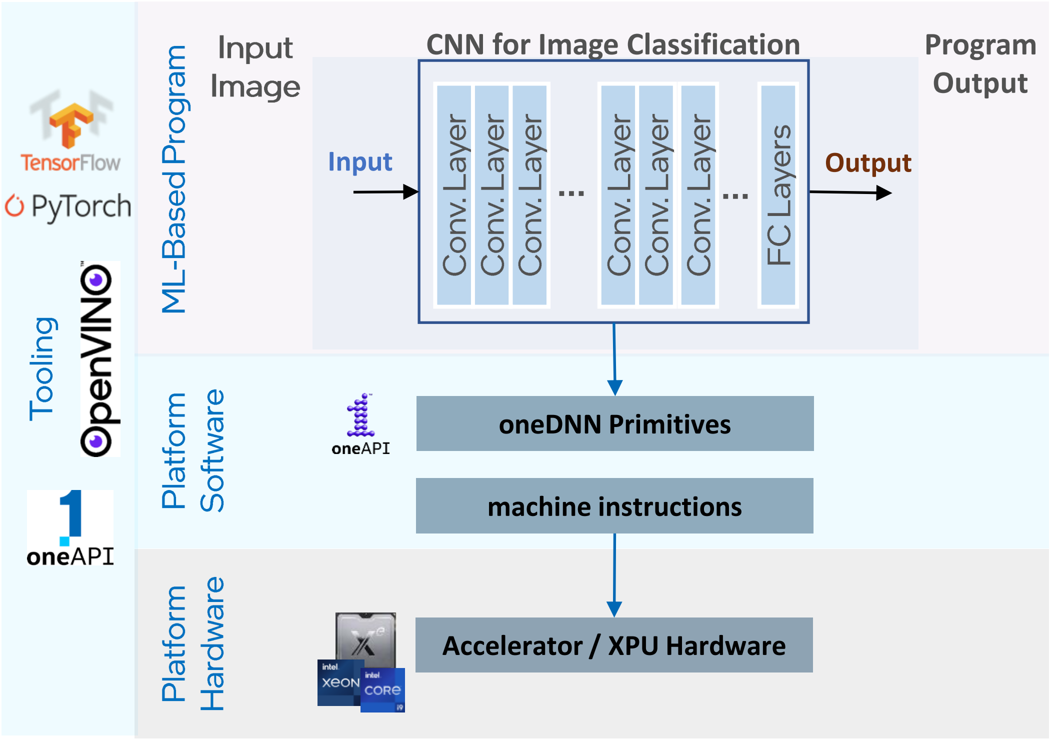
Error Impact Estimation Influences Hardware Protection

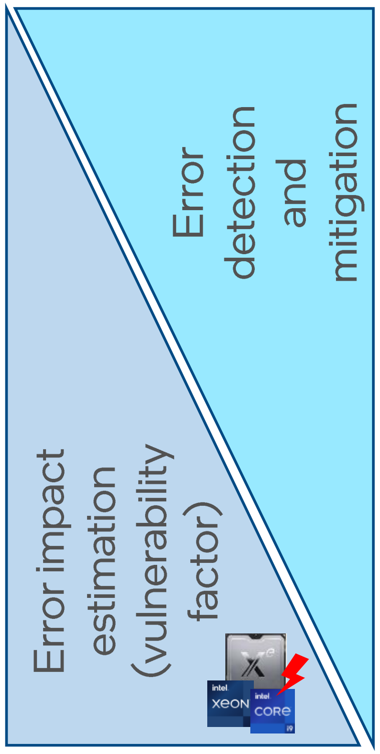
Dependable AI / ML - Resiliency

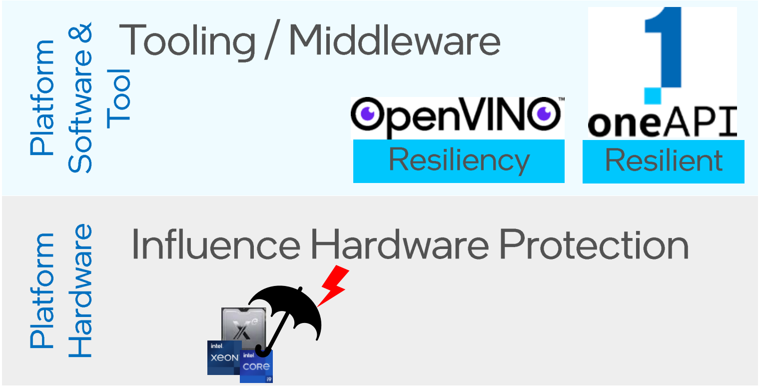
High-Performance Computing Safety-critical Applications

dependability / reliability target

Dependability Means







System View Dependability Threats Details on Dependability Means

Research Question: Mitigation Mechanisms at SW & Tooling Level

Dependable AI / ML - Resiliency

High-Performance Computing

Safety-critical Applications

dependability / reliability target

System View

Tooling: TensorFlow, PyTorch, OpenVINO, oneAPI

ML-Based Program

Input Image → Input → CNN for Image Classification [Conv. Layer | Conv. Layer | Conv. Layer | ... | Conv. Layer | Conv. Layer | Conv. Layer | ... | FC Layers] → Output → Program Output

Platform Software (oneAPI)

oneDNN Primitives

machine instructions

Platform Hardware

Accelerator / XPU Hardware

Dependability Means

Error detection and mitigation

Dependability Threats

Error impact estimation (vulnerability factor)

Details on Dependability Means

ML-Based Program

Robust DNN Architectures

example → Conv. Layer | Conv. Layer | ... | Resiliency Layer | ... | Conv. Layer | Conv. Layer | ... | FC Layers → ✓

Platform Software & Tool

Tooling / Middleware

OpenVINO Resiliency

oneAPI Resilient

Platform Hardware

Influence Hardware Protection

Research Question: Resilient Networks

Range Supervision

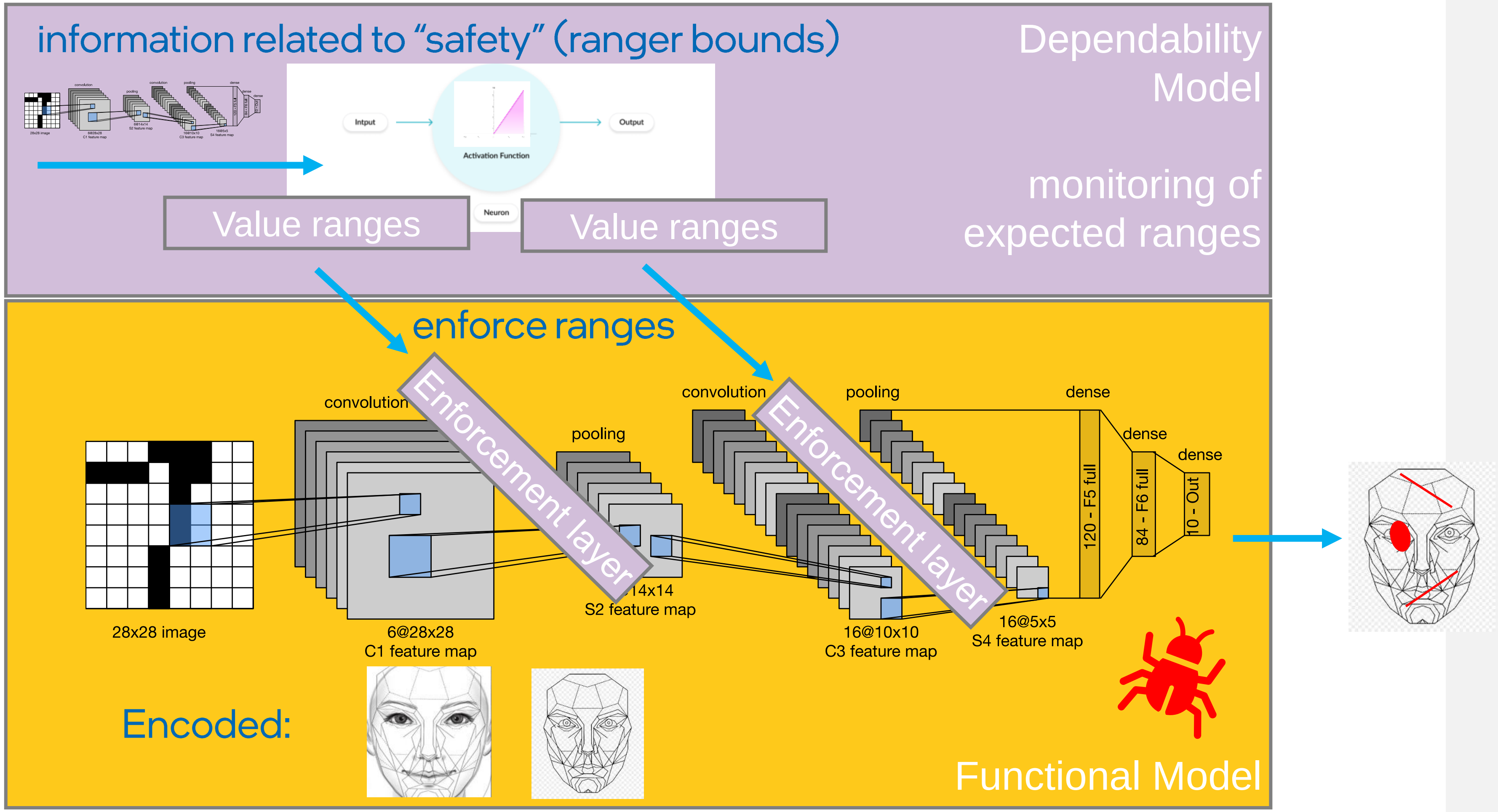
Example for a simple mitigation approach (joint work with Univ. of British Columbia – Prof. Karthik Pattabiraman)

“Independent” Additional Info?

Data

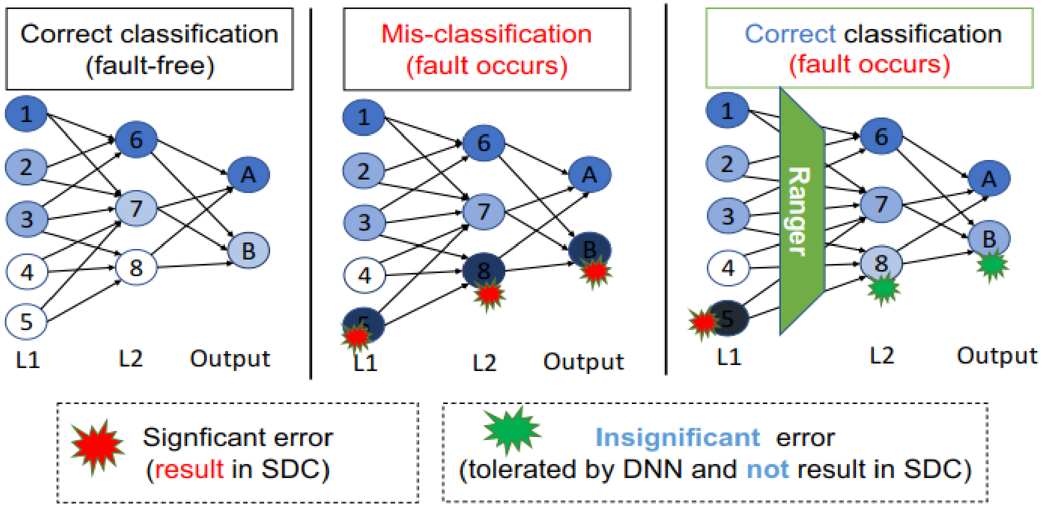
Used to extract activation value ranges

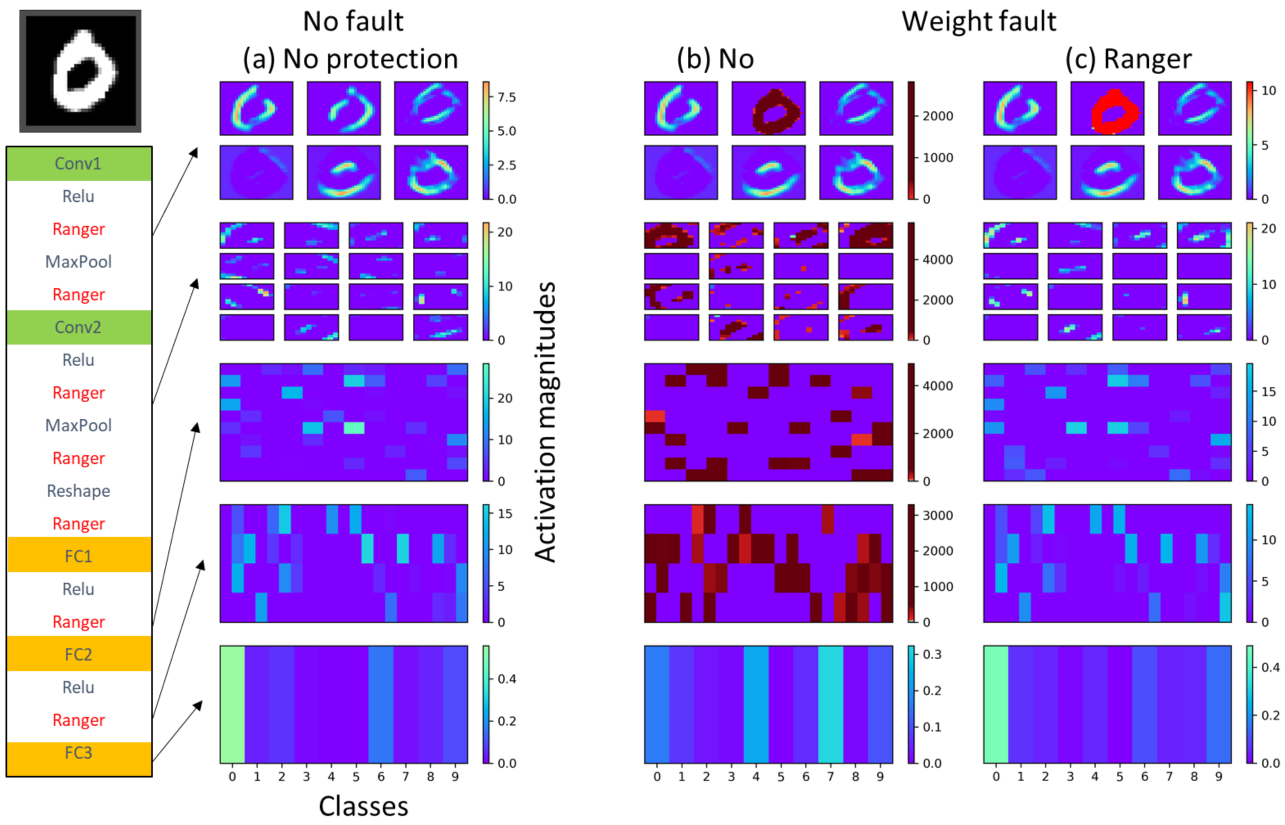
Used to train



Ranger – Intuition (I)

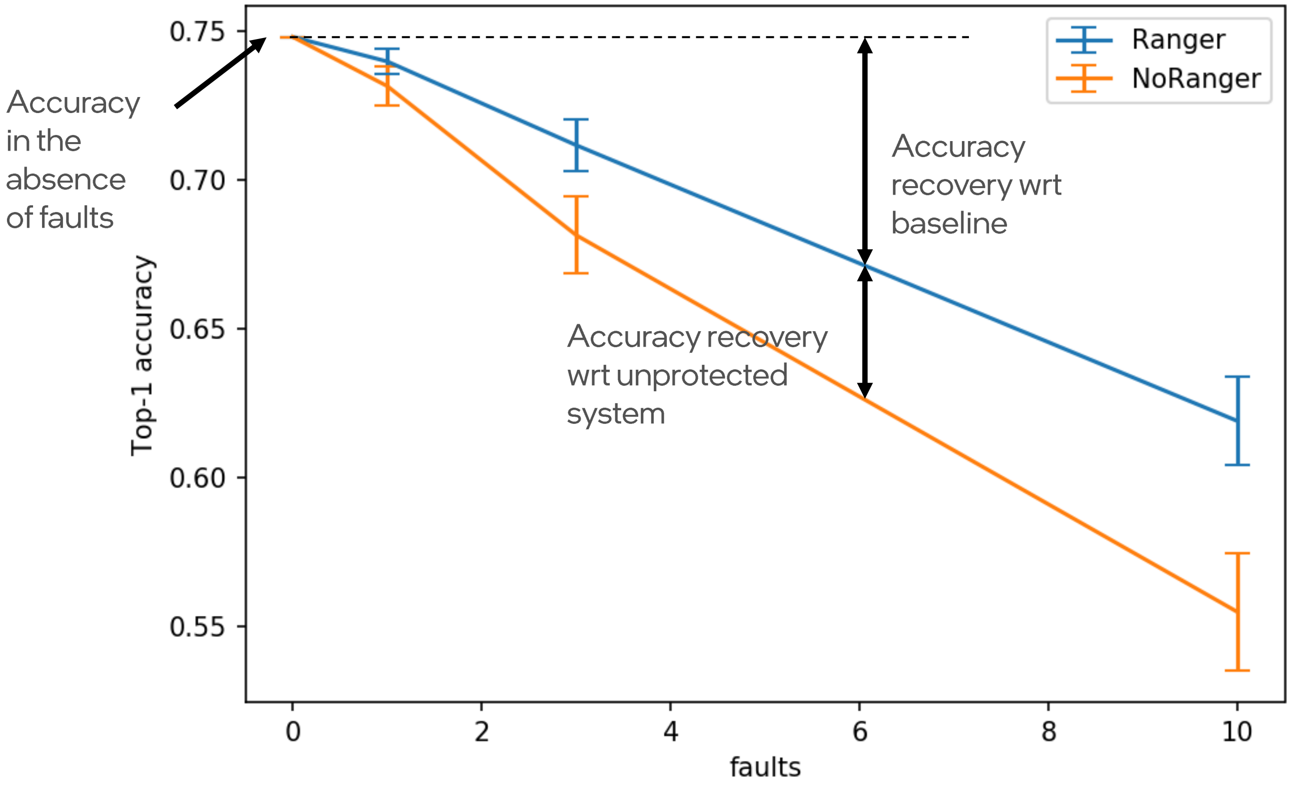
Chen et al 2020 (“Ranger”)
Li et al, 2017
Hong et al, 2019
Hoang et al, 2019 (“ClipAct”)



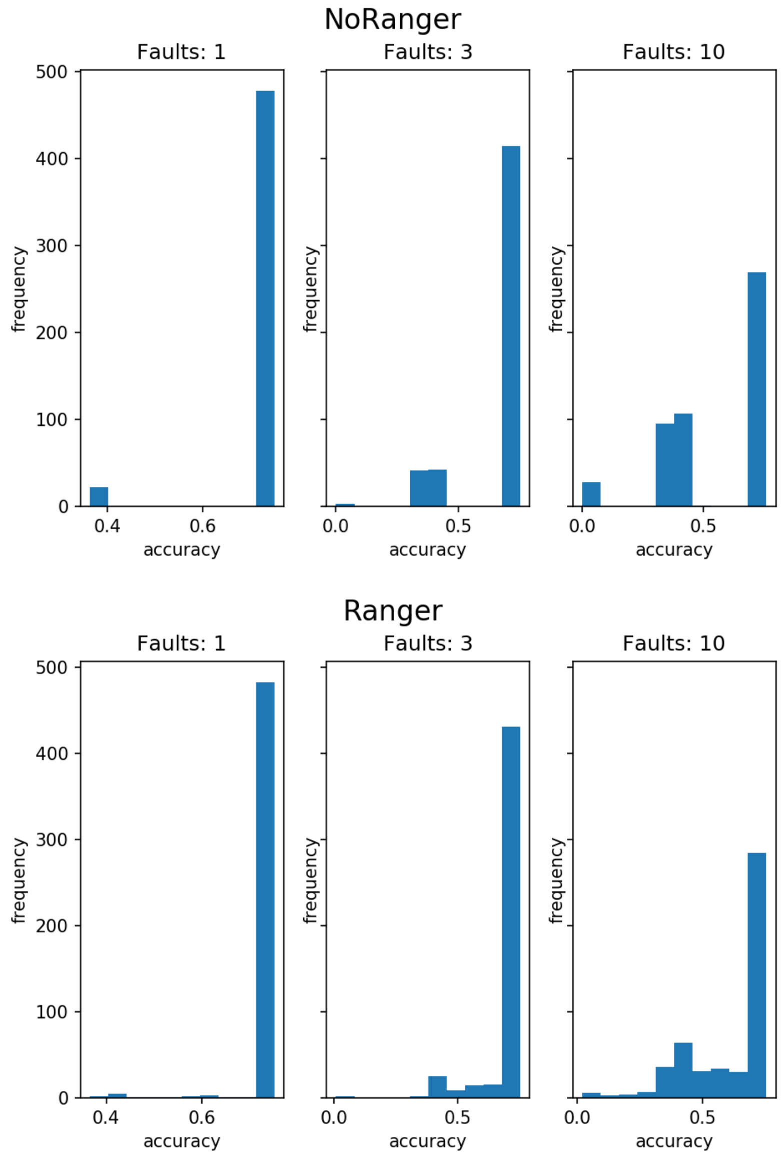


- Insert customizable protection layers (“Ranger layers”) for activation range restriction
- **Bound extraction** from an independent dataset (e.g. training data)
- No retraining of parameters needed

Ranger – Intuition (II)



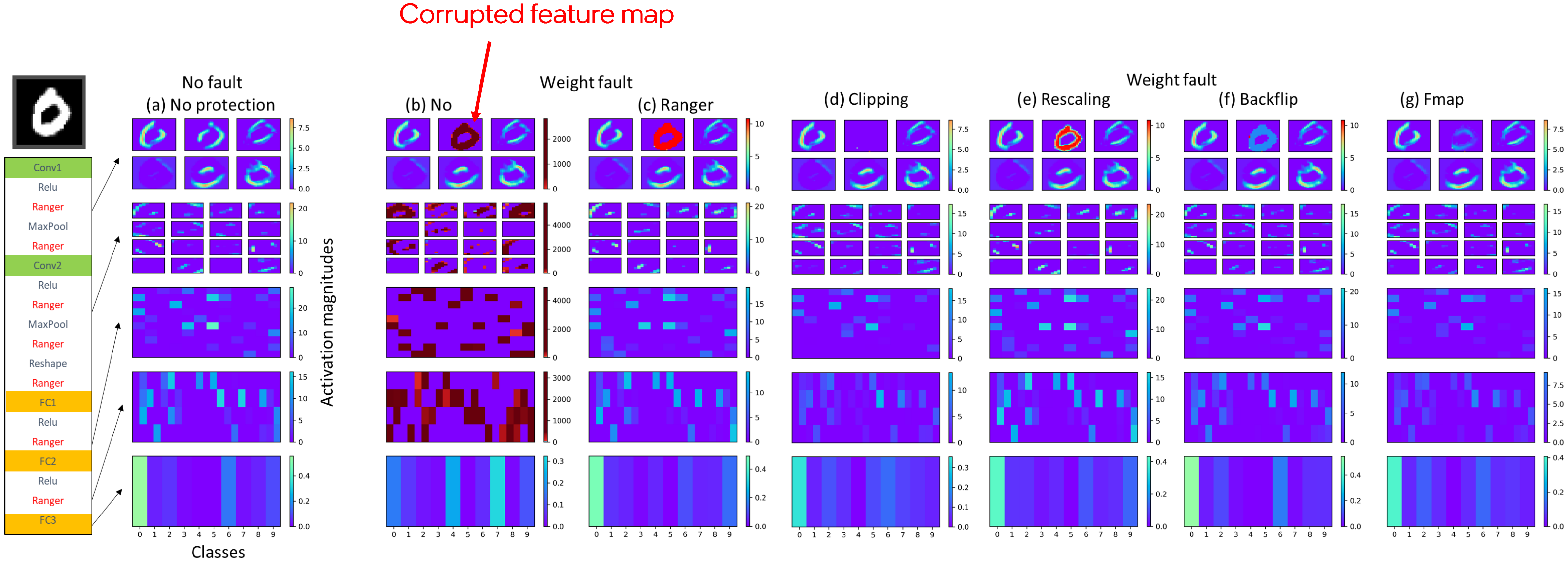
Distribution of accuracy results in 500 epochs:



- Large fault injection space: 1 weight fault in conv layers means fault rate of ~6.7e-8.
- Ranger mitigates the detrimental effect of faults per epoch by eliminating "outliers", and shift bulk towards maximum accuracy

Hoang et al, 2019 (“ClipAct”)
Geissler et al., 2021 (AISafety workshop, accepted)

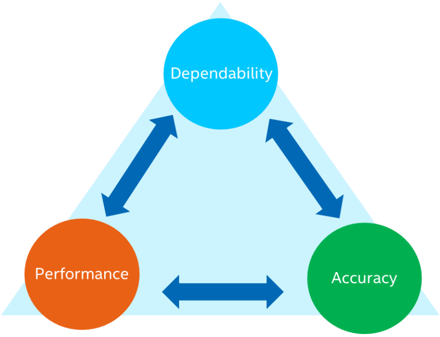
Range restriction alternatives



Goal is to restore the topology of feature maps after a soft error.

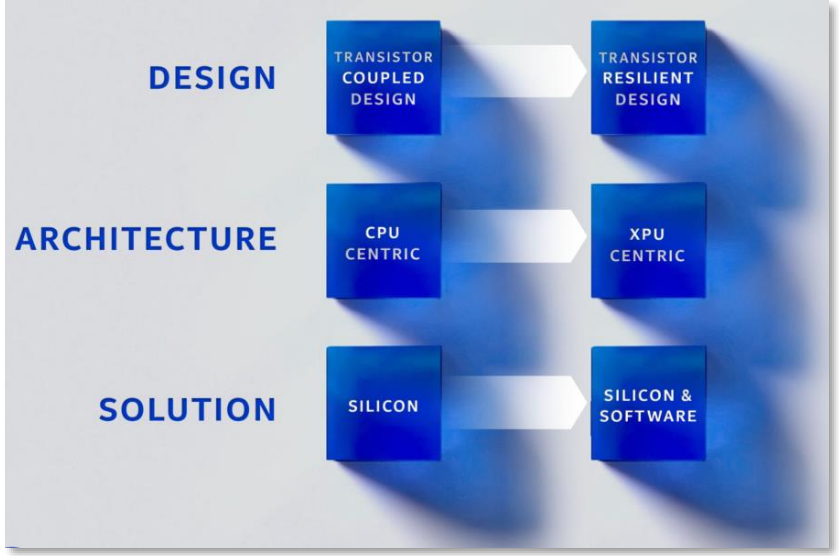
Some Ranger Findings

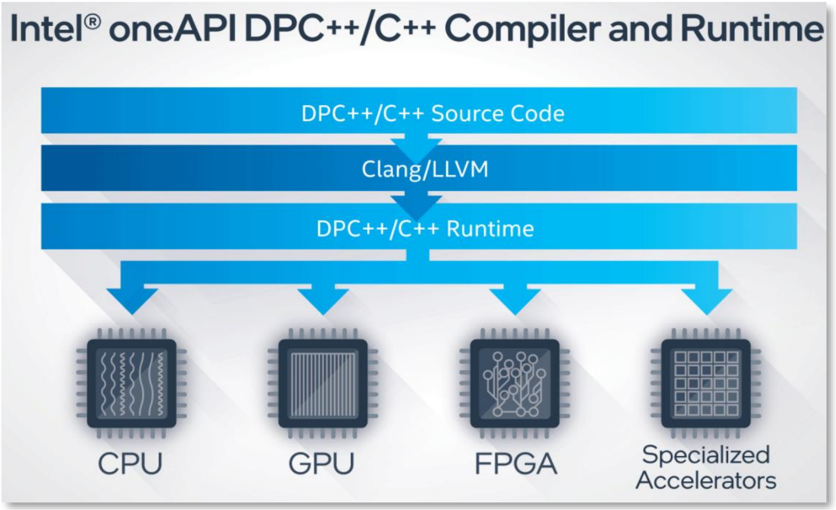
- If **bounds are extracted** from appropriate data, there is almost no reduction of the **baseline accuracy**.
- Impact depends on **data representation**: For FP32 and the given setup, almost all misclassifications happen due to flip in the MSB.
- Error detection: Strong correlation between out-of-bound events and misclassification. Range supervision provides **very high recall** (>0.99), **precision can be lower** (>0.82) due to false positives.
- **Error mitigation: Very high**, especially **Clipper/Backflip**. SDC rate is reduced by up to ~50x, to <0.5% in the studied setups.
- Some details @ Geissler et al. Towards a Safety Case for Hardware Fault Tolerance in Convolutional Neural Networks Using Activation Range Supervision, AISafety WS 2021
- Overhead: practically for free in practical scenarios



Summary & Outlook (Resiliency)

- **Dependability**: some faults are real (see also recent publications [1, 2])
- **Cost**: automation and low overhead key to acceptance
- **Software and tools** play a larger role these days
  - Open source and open languages
  - Libraries oneDNN / oneAPI
  - Cross-architecture languages, compilers, and tools
    - DPC++ = ISO C++ and Khronos SYCL™ and community extensions
    - OpenVINO™ tool
  - Researchers can engage in open source push





Special thanks goes to DRL team and Karthik Pattabiraman

[1] https://sigops.org/s/conferences/hotos/2021/papers/hotos21-s01-hochschild.pdf [2] https://arxiv.org/abs/2102.11245

Monitoring for Safe Perception

Application and System Level

Safe Perception - Monitoring

- Perception in complex environment with **multi-level approaches** to improve safety:
- Application-level context (view angles, infrastructure involvement, …)
- System-level context (diverse space and object representation, sensors, monitors, …)

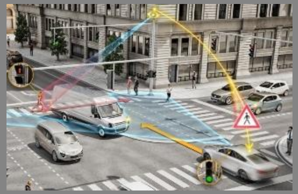
Largest Benefits from Infrastructure

VS. single automated car

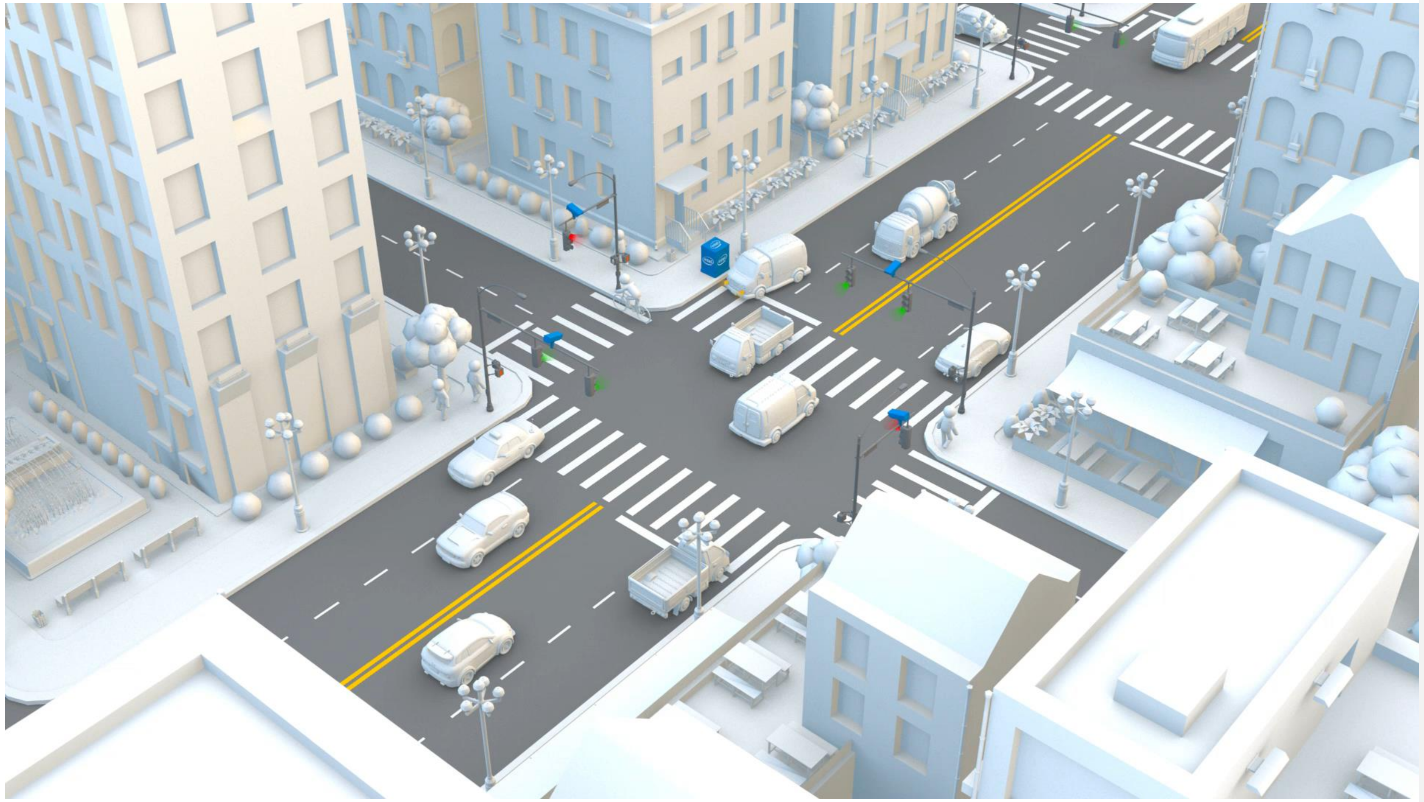
Single car

vs.

Infrastructure



- Additional independent source of perception
- Extended field of view
- Different perception vantage points
- Extended compute and energy envelope

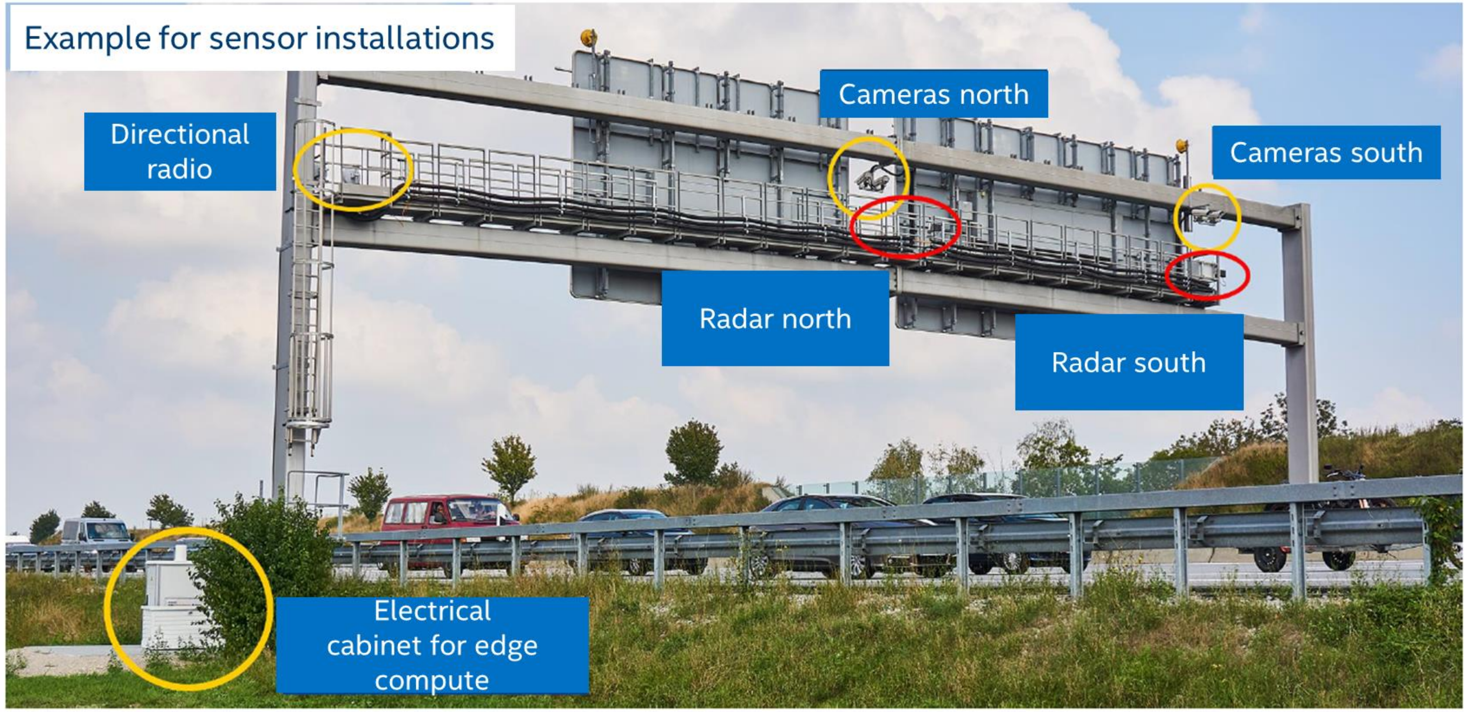


https://www.geospatialworld.net/news/mobileye-join-hands-enable-crowd-sourced-hd-mapping-automated-driving/
https://wtvox.com/fashion-innovation/the-future-of-driving-with-v2v-and-v2i-technology/

26

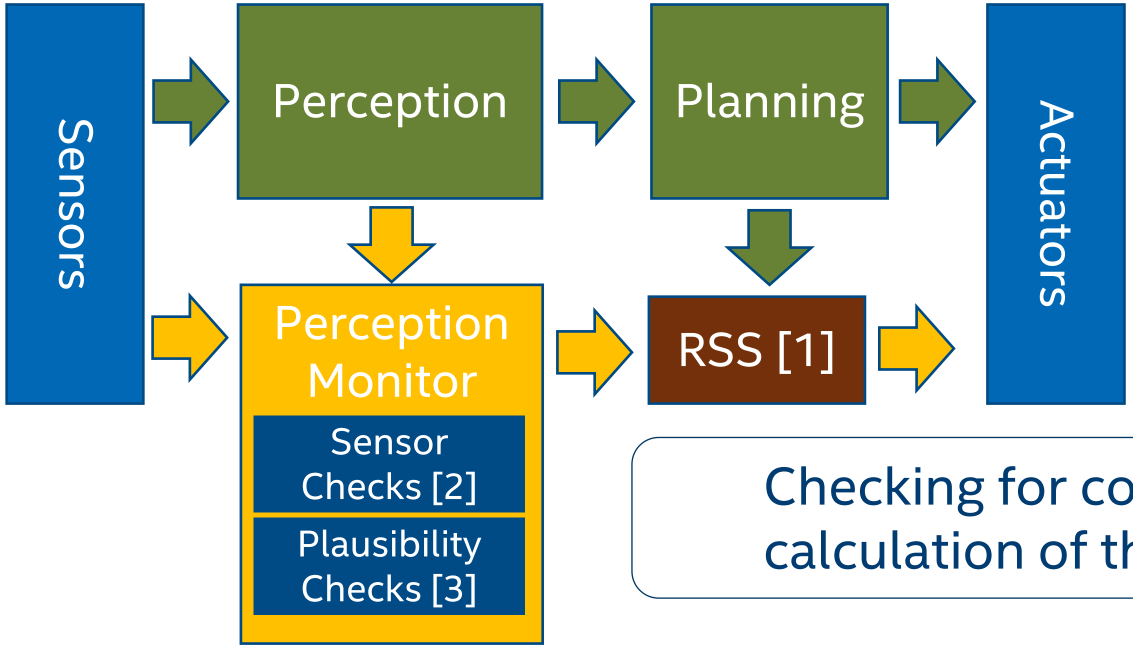
Providentia++

- Basis of the digitalized highway of the future: Real-Time Digital Twin for Smart Highway & Smart City
- Benefits:
  - Real life use case for dependability work
  - Test bench for work of Intel Labs (ASMRL & DRL teams)
  - Work with IOTG Autonomous Transportation and Infrastructure

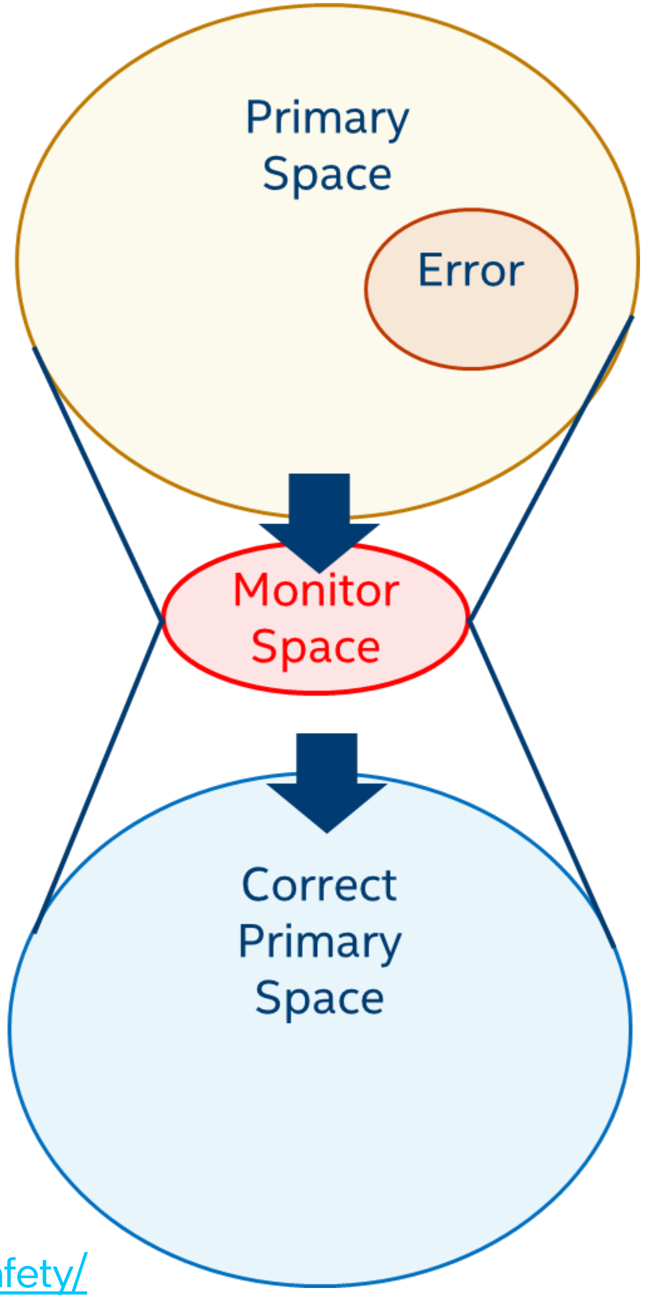


Monitor Architecture at Board Level with Application and Application Monitor

How to implement a system that can monitor & recover function :
- requiring significant **less complexity**
- **not decreasing availability** of the primary channel

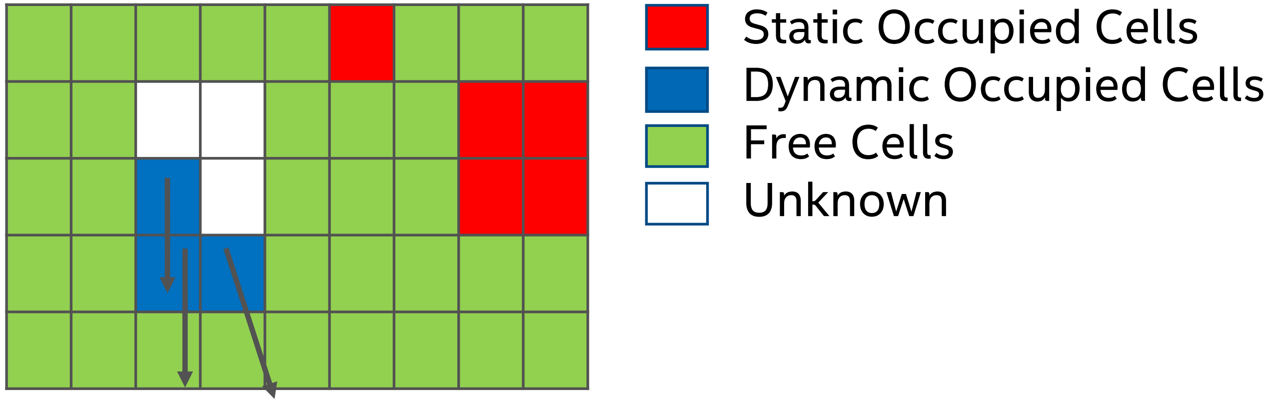


Checking for correctness easier than calculation of the correct result

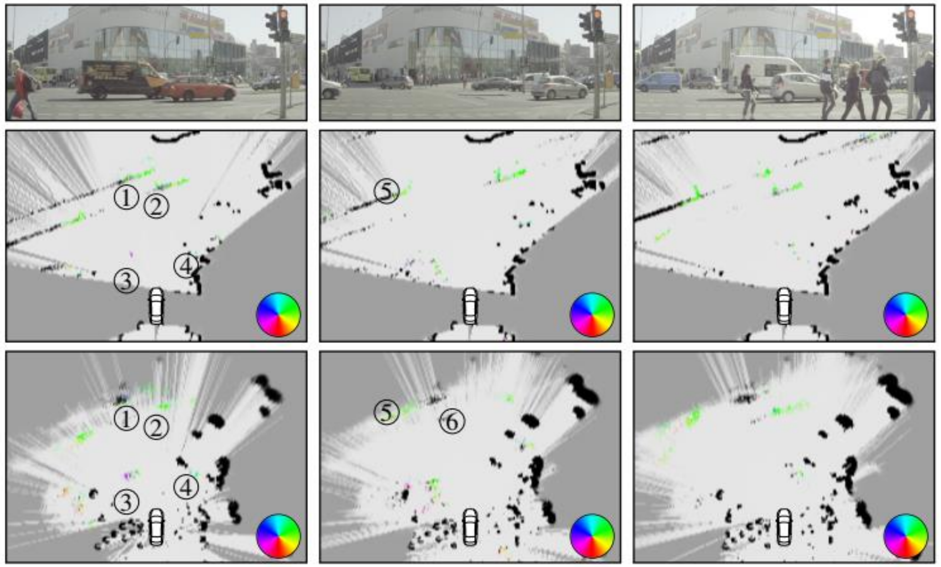


[1] https://www.mobileye.com/responsibility-sensitive-safety/
[2] https://ieeexplore.ieee.org/iel7/9304518/9304528/09304571.pdf
[3] https://arxiv.org/pdf/2009.14756

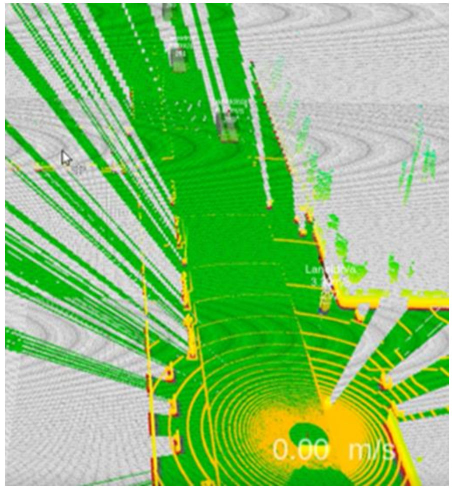
Dynamic Occupancy Grid



[3]



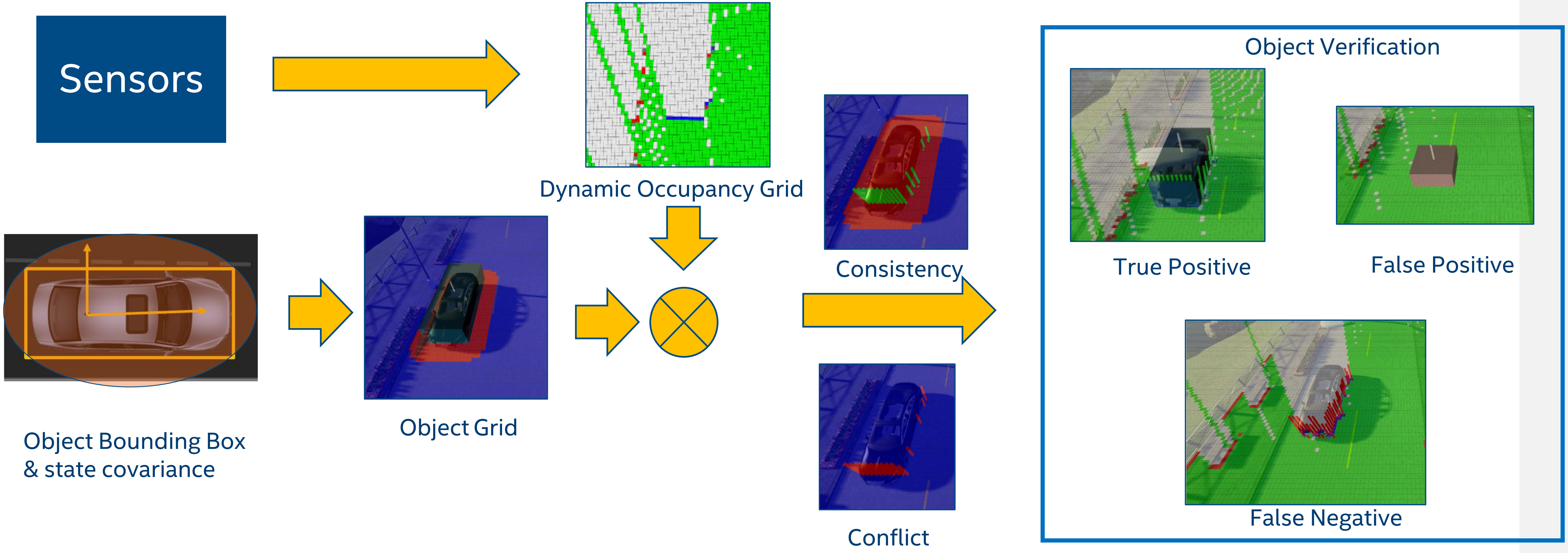
- Promising results in cluttered & dynamic environments
- Fusion of Lidar [1][2] and Radar sensor[3] information
- *Classical algorithm* – redundancy towards ML algorithms



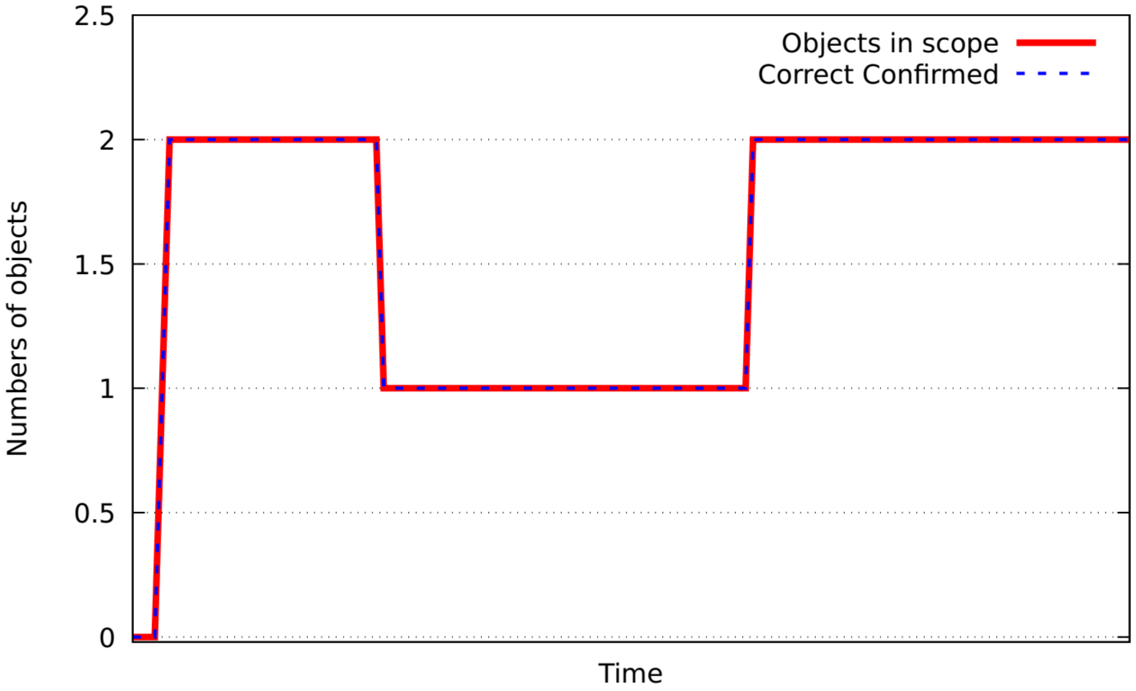
[1] G. Tanzmeister and D. Wollherr, “Evidential Grid-Based Tracking and Mapping,”
[2] D. Nuss, et al., “A random finite set approach for dynamic occupancy grid maps with real-time application”
[3] Christopher Diehl, et al. “Radar-based Dynamic Occupancy Grid Mapping and Object Detection”, ITSC 2020

Sensor Checks - Position

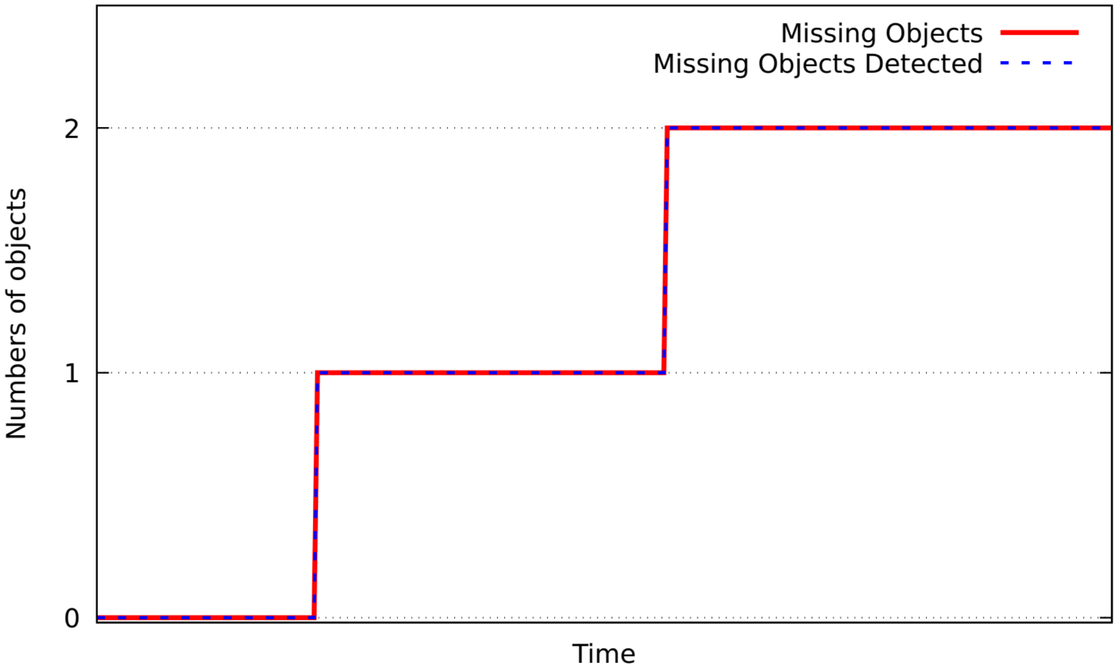
Objects
- **Position**
- Velocity
- **Dimension**



Results: True Positives & False Negatives



True Positives



False Negatives

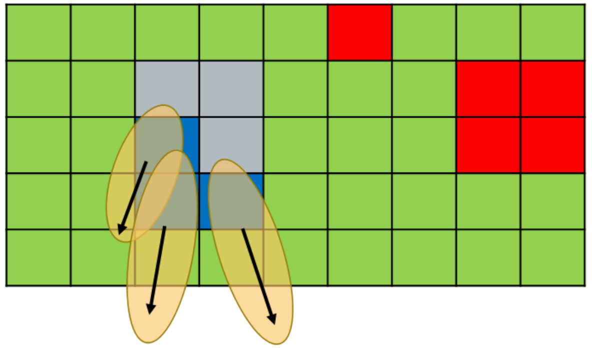
ITSC2020 – *Towards Online Environment Model Verification* | Cornelius Buerkle, Fabian Oboril & Kay-Ulrich Scholl

Objects
- Position
- **Velocity**
- Dimension

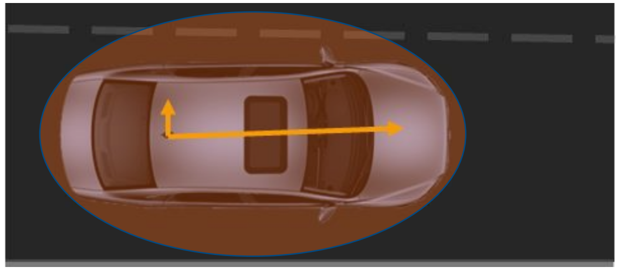
Velocity checks

Sensor checks

Calculate similarity / distance of velocity distribution of each cell in object region with velocity distribution of object

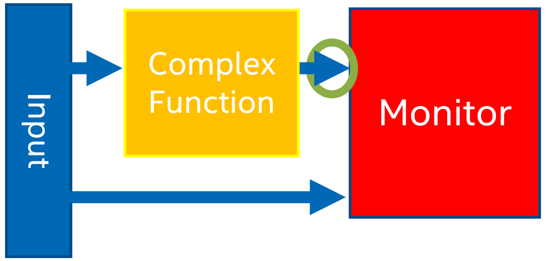


occupancy grid velocity information $V(\zeta)$

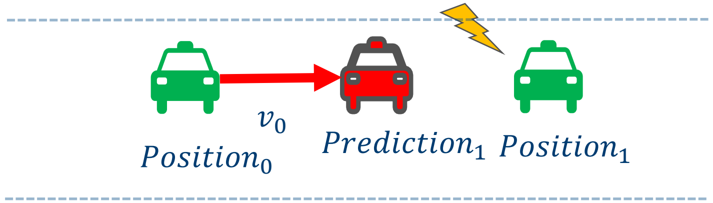


object velocity $\Sigma_v(o)$

Plausibility checks



Correct Velocity

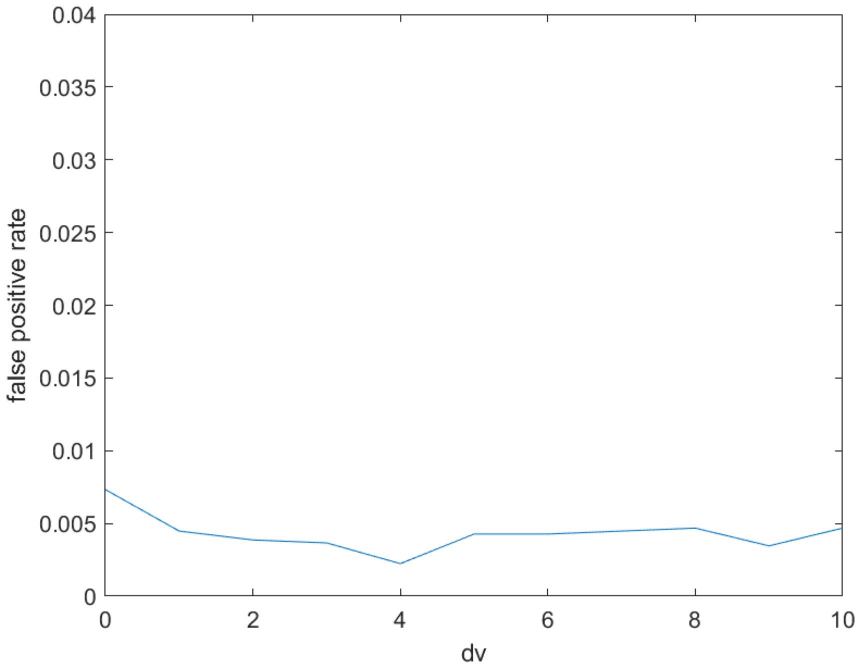


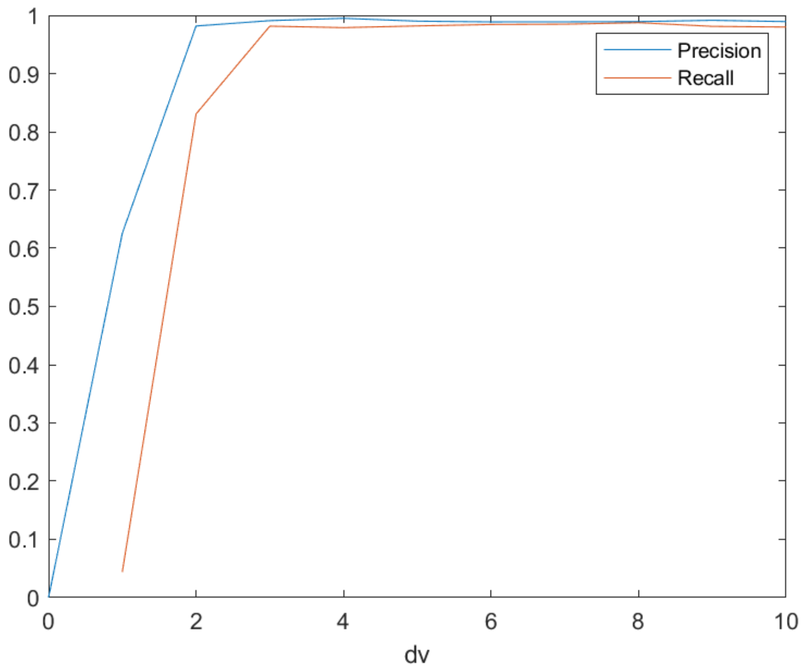
False Velocity

https://arxiv.org/abs/2009.14756

Results – Plausibility checks velocity

Chance of error injection per vehicle and time step: **0.25**
Scenario time: Nuscenes clip of ~**125** time steps
Time intervals: ~**0.2s**





- False positive rate (FPR) in absence of faults **~ 0.5%**
- Measured against **ground truth,** where avg velocity information estimates were added (pos diff over time from last time step).
- Rather insensitive to speed error.

- Increasing speed error faults dv injected. Here, dv = 0 means no faults injected
- **Recall ~ 0.98 and Precision ~ 0.98 with the given parameters for dv ≥ 3m/s (or ~11km/h)**

Addressing Safety at Different Levels …

Main lane
Main function
Main value

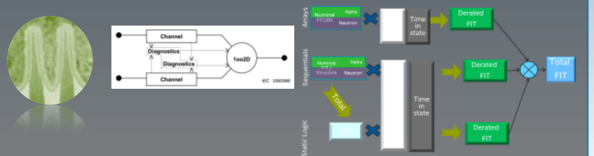
Monitor lane
Safety function
Safety value

CHIP LEVEL

Powerful computing e.g.
to master perception challenge

High integrity and availability requires simple approaches

Reliability | Integrity | Robustness



BOARD LEVEL

Powerful Computing and Accelerator

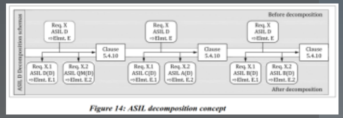
intel XEON PLATINUM inside | intel NERVANA inside | intel STRATIX inside

Safety Monitoring (Doer/Checker) for systematic and random hardware faults

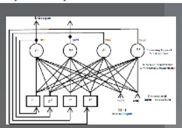
Monitor

intel ATOM inside

ASIL decomposition

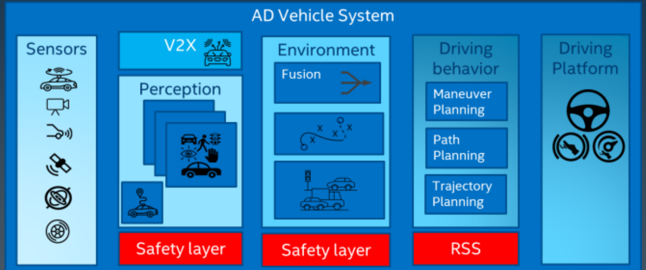


Intelligent diagnostic
& monitoring



AD APPLICATION LEVEL (SYSTEM LEVEL)

Complex AD Stack (with AI like DNN in perception, powerful multi-level)



Simple Checker App. –
e.g. Responsibility Sensitive Safety (RSS)

… can sometimes be synergetic

Summary (Monitoring)

- Monitoring at different levels helps dependability (i.e. safety in this context)
- Application-Level Monitor approaches could help to solve the safety challenge for automated vehicles (AVs)
  - Showed initial realization for monitors of object information based on sensor and plausibility checks
  - Demonstrated feasibility with evaluation results in simulation
- Challenges:
  - Sensor checks rely on quality of dynamic occupancy grid and sensor preprocessing
  - Erroneous cells occupation decrease availability or error detection
  - Diversity arguments and proofs for effectiveness
  - Error detection effectiveness to be further investigated (incl. latent errors)