FUNCTIONAL SAFETY AND THE GPU

Richard Bramley, 5/11/2017

NVIDIA

AGENDA

How good is good enough

What is functional safety

Functional safety and the GPU

Safety support in Nvidia GPU

Conclusions

HOW GOOD IS GOOD ENOUGH ?

ACCIDENT STATISTICS- US[1]

Description	2013 Statistics	2015 Statistics
Fatal Crashes	30,057	35,092
Non-Fatal Crashes	5,657,000	6,263,834
Number of Registered Vehicles	269,294,000	281,312,446
Licensed Drivers	212,160,000	218,084,465
Vehicle Miles Travelled	2,988,000,000,000	3,095,373,000,000
Fatal Crash Rate in FITs [2,3]	250 - 500	283 - 566
Non-Fatal Crash Rate in FITs [2,3]	46K - 92K	51K - 102K
What is an appropriate target ?		

Google Non-Fatal Crash FIT Rate = 150K

[1]Source: Traffic Safety Facts 2013/2015, NHTSA document reference DOT HS 812384
[2] Derived from NHTSA data on driver related fatal crashes
[3]Assumes an average speed of 50MPH

TARGET FAILURE RATES

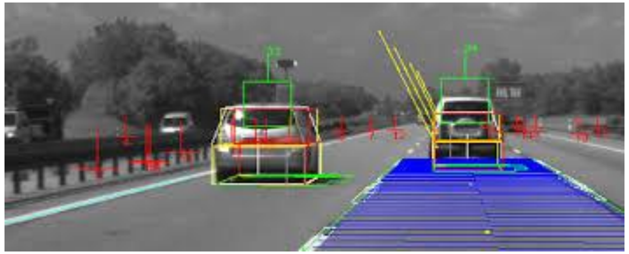
Description	Statistics
Acceptable risk (no further improvement required)	1:1,000,000[1]
US population (2015)	>321,000,000
Traffic deaths	35,092
“Acceptable” deaths as per guidelines	321
Required improvement	x100

Wide variety of targets in industry
Target risk reduction of 2x to 100x compared to human driver

[1] Derived following data from UK health and safety executive publications

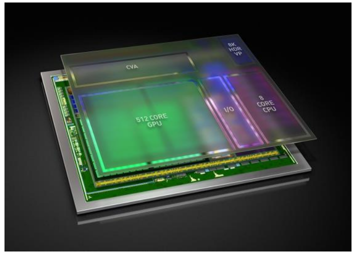
SAFETY AND AUTONOMOUS VEHICLES

Algorithms



Software

Hardware



Safety during intended operation
Safety of the intended function
(SOTIF ISO/PAS 21448 in development)

Safety in presence of a fault
Functional Safety ISO-26262

FUNCTIONAL SAFETY BASICS

DEFINITION PER STANDARDS

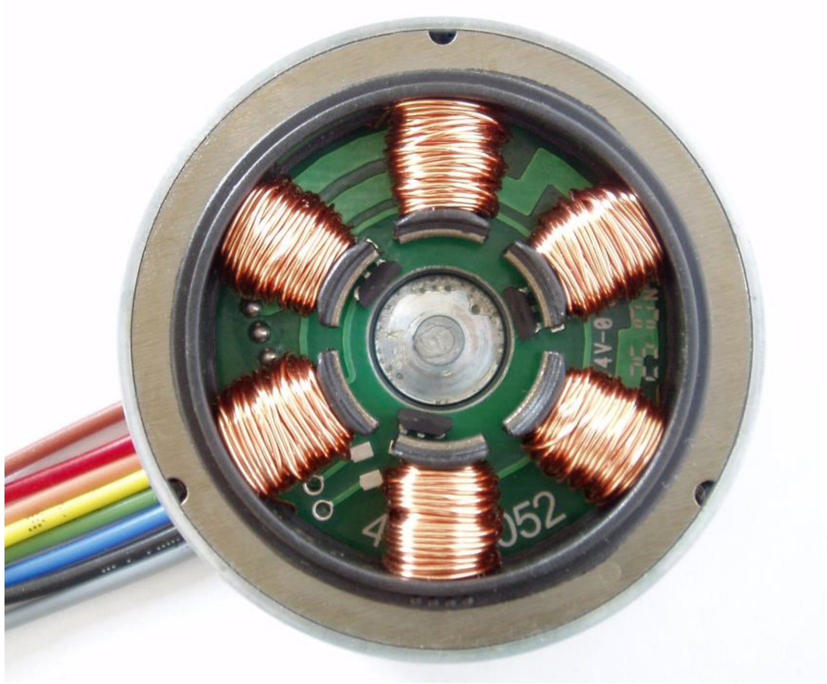
“*Absence of unreasonable risk due to hazards caused by malfunctioning behavior of electrical/electronic systems*” – ISO 26262-1:2011; 1.51

“*Part of overall safety relating to the equipment under control and the equipment under control, control system that depends on the correct functioning of the electrical/electronic/programmable electronic safety-related systems and other risk reduction measures*” – IEC 61508-4:2010; 3.1.12

CLASSIC EXAMPLE

IEC 61508-0:2005; 3.1

- Consider a motor winding which may overheat and cause a hazard.
- Reliability engineering approach might design the winding to be more resilient to over-temperature conditions
- Functional safety engineering approach might add a temp sensor to detect the over-temperature condition and switch off the motor



https://upload.wikimedia.org/wikipedia/commons/0/0f/Stator_Winding_of_a_BLDC_Motor.jpg

ACHIEVING FUNCTIONAL SAFETY

Systematic and random faults must be considered

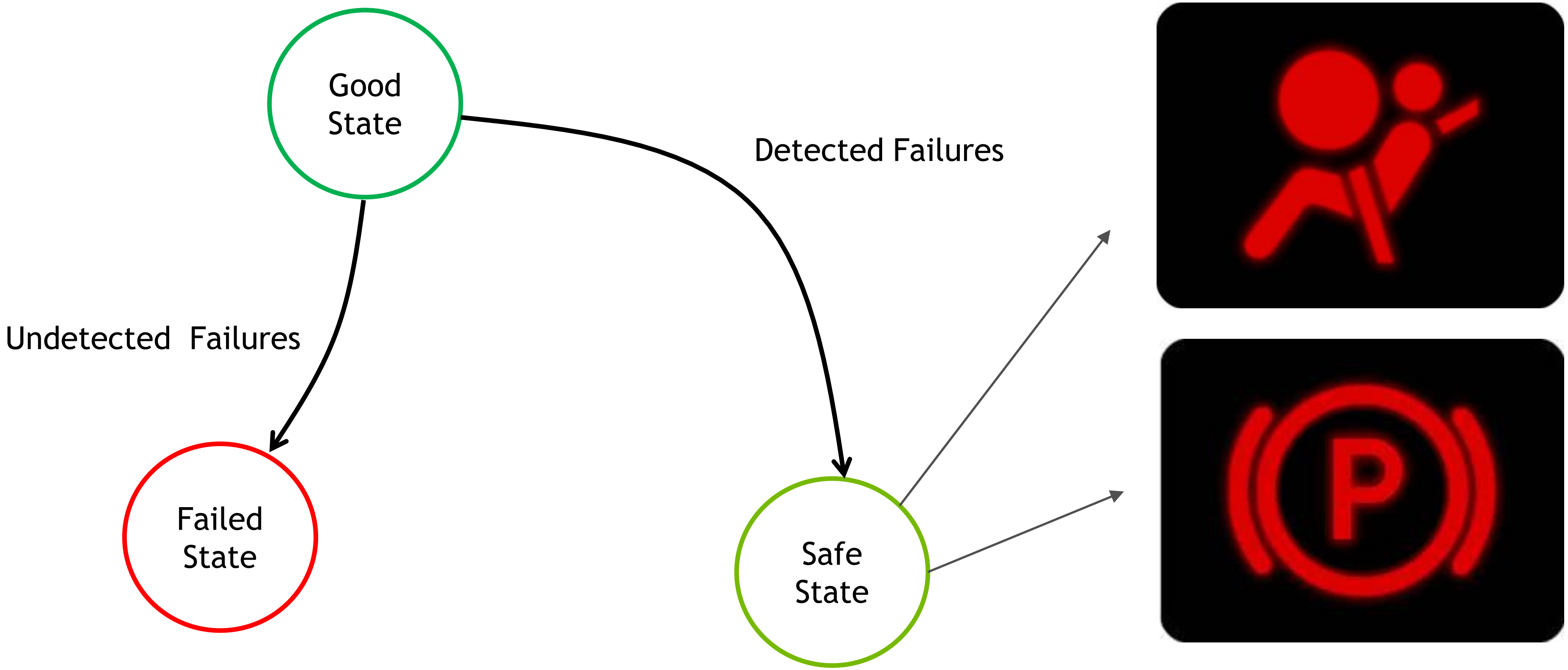
Systematic faults mitigated by:

- Following compliant process at all stages of development
- Monitoring of the complete product lifecycle

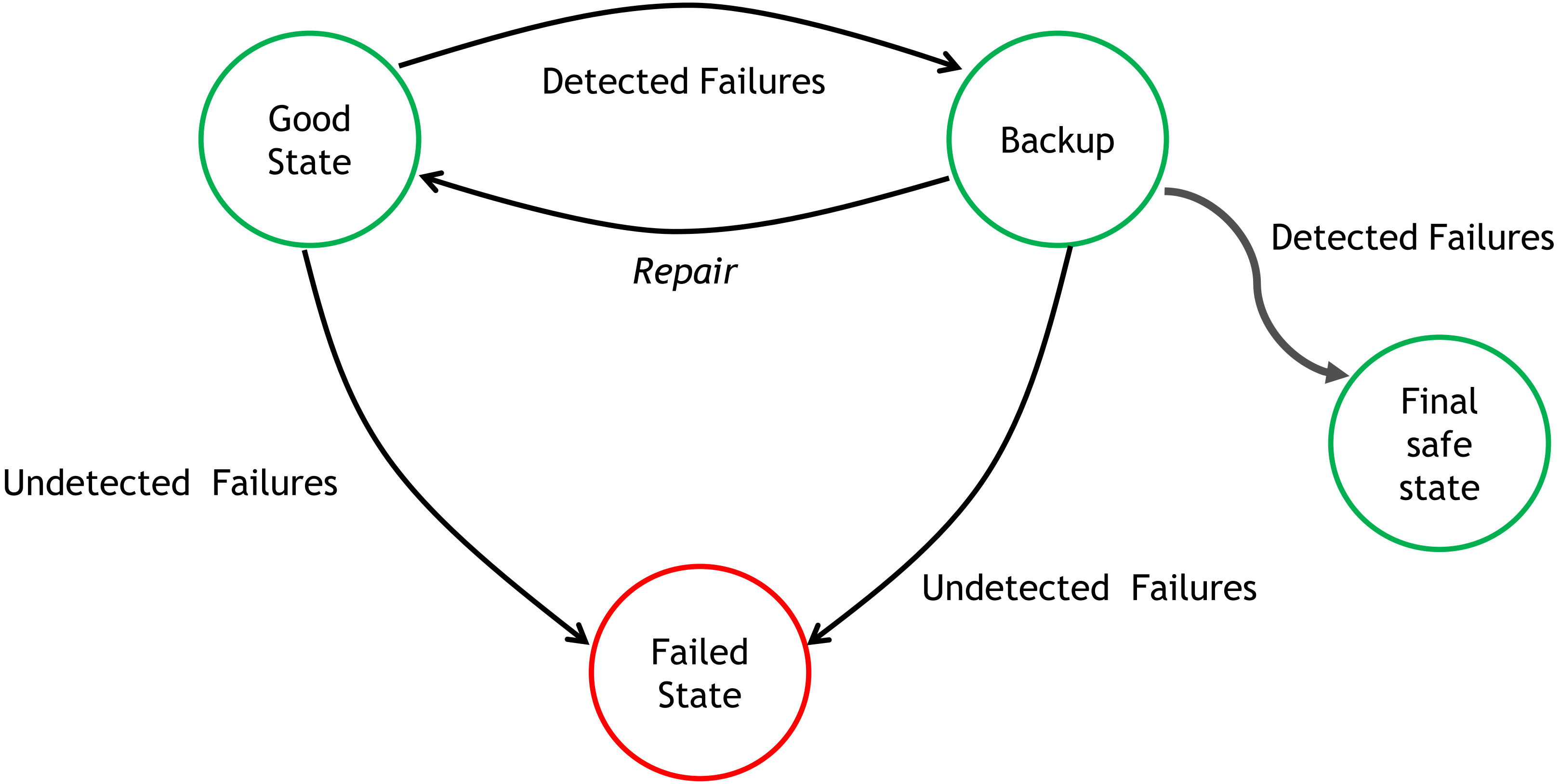
Random faults are mitigated by:

- Failure mode analysis to understand the fault behavior of the system
- Application of diagnostic measures to detect the failure modes
- Transition to the safe state on failure mode detection

FAIL SAFE


Good State
Detected Failures
Undetected Failures
Failed State
Safe State
P

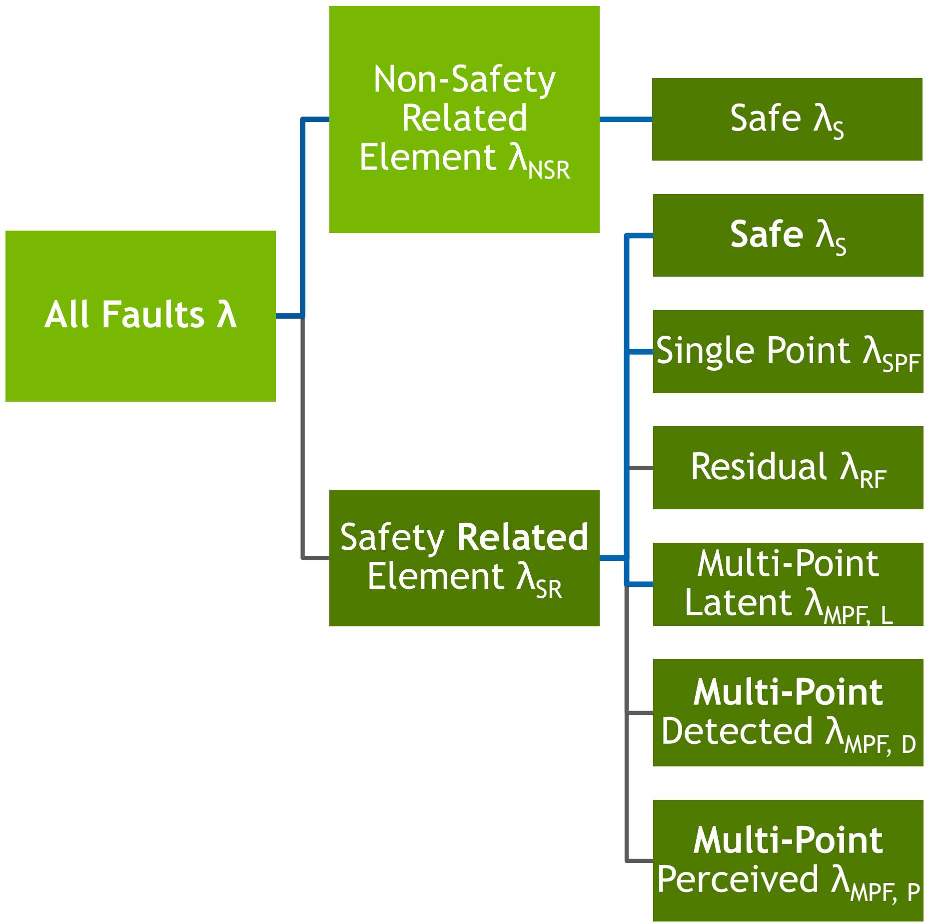
FAIL OPERATIONAL



For full autonomy the initial “safe state” can be a transition to the backup system

FAULT CLASSIFICATIONS

ISO 26262-10; B.1



SINGLE POINT FAULT METRIC (SPFM)

Shows the percentage of overall single point faults which are:

Safety related AND

Safe OR dangerous but detected

$$1 - \frac{\sum_{SR,HW} (\lambda_{SPF} + \lambda_{RF})}{\sum_{SR,HW} \lambda} = \frac{\sum_{SR,HW} (\lambda_{MPF} + \lambda_{S})}{\sum_{SR,HW} \lambda}$$

λ_s - safe fault failure rate, can also be expressed as a % (Fsafe) the ration of overall possible faults which are safe.

LATENT FAULT METRIC (LFM)

Shows the percentage of overall multiple point faults which are:

- Safety related AND
- Safe OR dangerous but detected OR dangerous but perceived

Customarily limited to scenarios considering 2 point independent faults

Primary consideration is fault in mission logic combined with fault in safety mechanism

$$1-\frac{\sum_{SR,HW}(\lambda_{MPF,latent})}{\sum_{SR,HW}(\lambda-\lambda_{SPF}-\lambda_{RF})}=\frac{\sum_{SR,HW}(\lambda_{MPF,perceived\ or\ detected}+\lambda_{S})}{\sum_{SR,HW}(\lambda-\lambda_{SPF}-\lambda_{RF})}$$

ARCHITECTURAL METRIC TARGETS

	ASIL A	ASIL B	ASIL C	ASIL D
SPFM	N/A	>=90%	>=97%	>=99%
LFM	N/A	>=60%	>=80%	>=90%

All targets are recommendations. Developers can set their own targets based on appropriate argumentation.

PROBABILISTIC METRICS

Probabilistic Metric for (Random) Hardware Failure (PMHF)

Examines the residual probability of violation of safety goal after application of diagnostics, in a given time of operation.

$$M_{\text{PMHF}} = \lambda_{\text{RF}} + \lambda_{\text{m,DPF}} \times \lambda_{\text{sm,DPF,latent}} \times T_{\text{Lifetime}}$$

ISO 26262-10:2011; 8.3.3

Some pushback in market due to inconsistency between methods used by different vendors.

NOTE: Multiple versions of equation possible depending on conditional probability of failures. Simplest form shown

PMHF TARGETS

	ASIL A	ASIL B	ASIL C	ASIL D
PMHF	N/A	100 FIT	100 FIT	10 FIT

All targets are recommendations. Developers can set their own targets based on appropriate argumentation.

RELEVANCE TO GPU

EXAMPLES OF SAFETY CRITICAL OPERATION ON GPU

TRADITIONAL CV

Normalize gamma and color

Compute gradients

Weighted voting

Contrast and normalize

Collect HOGS

Traditional Classification: (pattern and template matching)

MACHINE LEARNING*

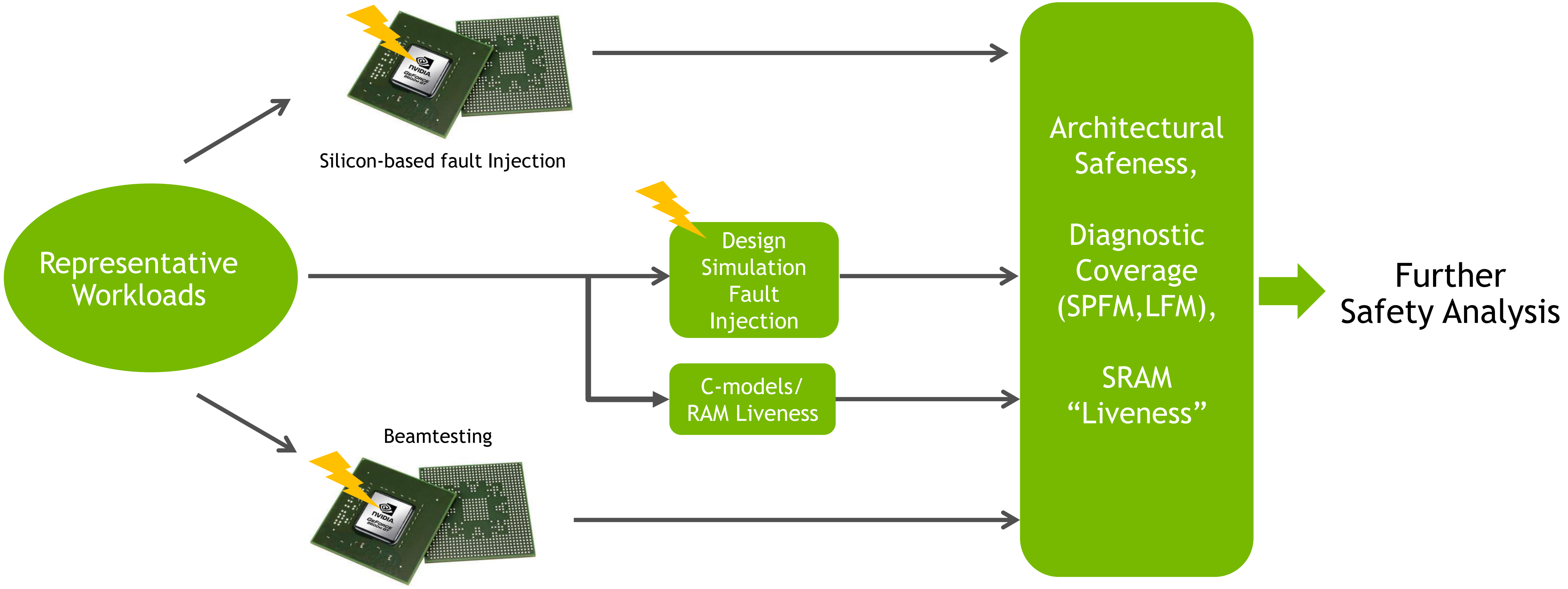
CNN (Convolutional Neural network)

MLP (Multi-layer perceptron)

SVM (Support vector machine)

*Focus is inferencing, training handled analogously to validation and calibration of a traditional safety related algorithm.

GPU MEASUREMENT METHODOLOGIES



Much of the measurement is done on representative kernels as the final applications are not available at design time

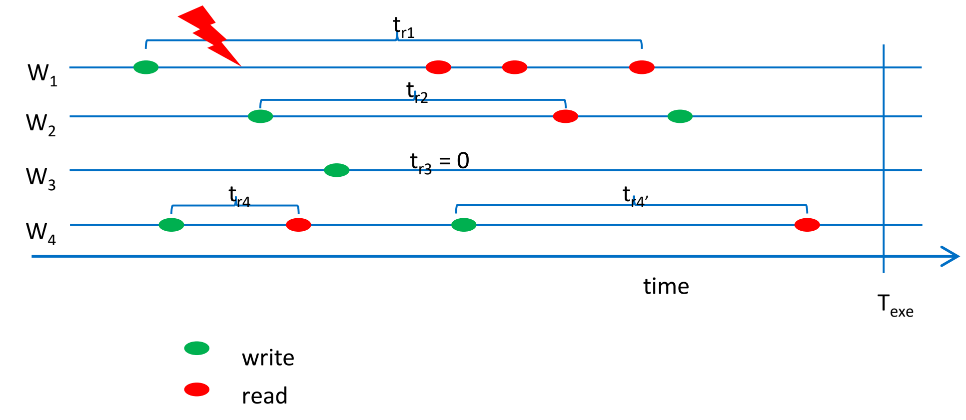
MEASURING SAFE FAULTS IN RAMS
"LIVENESS"

RAMs are sensitive to particle radiation (4x larger failure rate per bit than flops)

RAM contents may not be sensitive to faults (pixels)

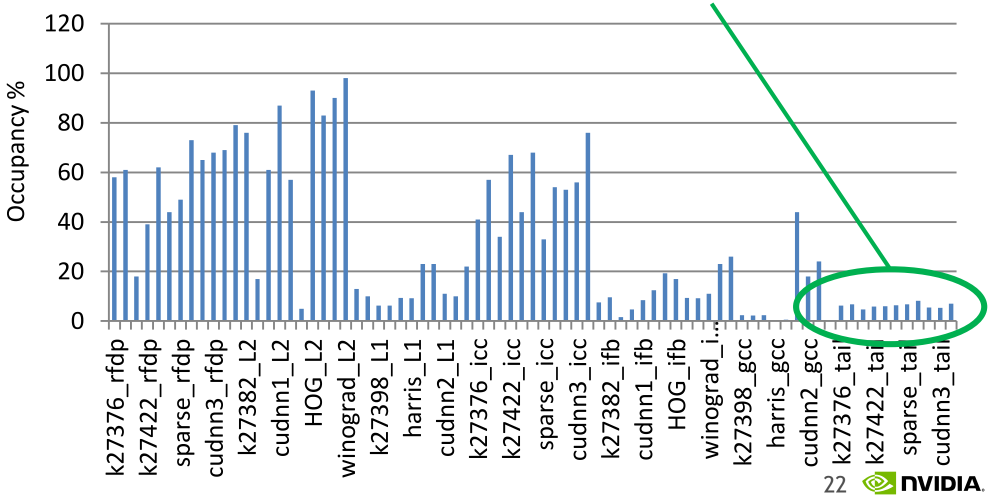
RAM contents may be very sensitive to faults (instructions)

An important indicator is RAM Liveness



The occupancy can be computed: $(t_{r1} + t_{r2} + t_{r3} + t_{r4} + t_{r4'}) / 4 \times T_{exe}$.

Majority of RAMS in this GPU less than 10% occupancy
Fsafe > 90%



TESTING REPRESENTATIVE KERNELS

Parameter measurement is very sensitive to kernel definition

Traditional CV has a wide diversity of operations

Difficult to define representative kernels

Machine learning has a smaller set of repeated operations

Enabling a more complete definition of kernels for measurements

More accurate and reliable measurements

DEEP LEARNING APPLICATION SAFENESS

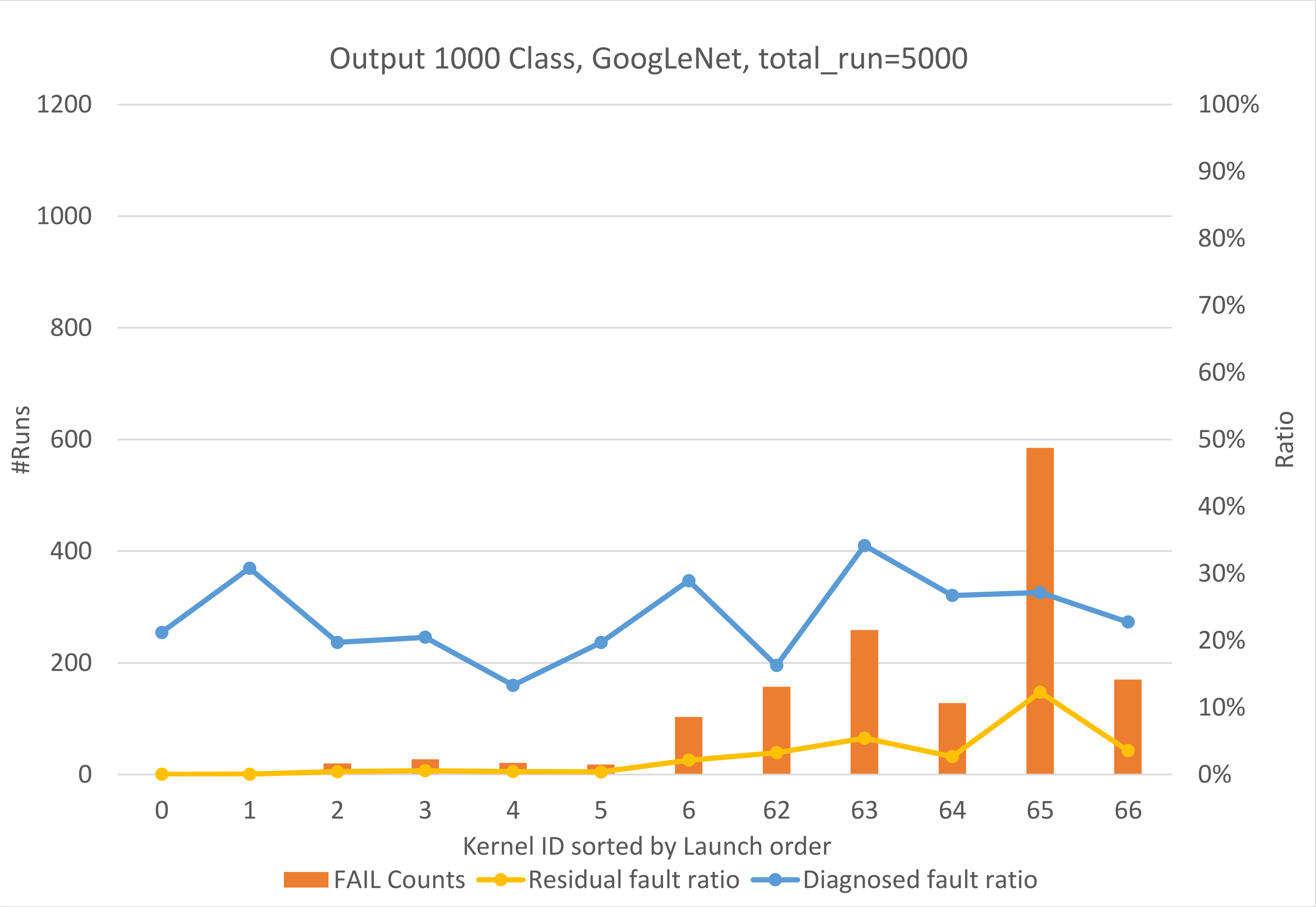
GIE GoogLeNet

67 kernels in GoogLeNet inference

Faults in latter kernels have a higher possibility to cause errors

#FAIL Counts represents the proportion of faults for which the application predicted the wrong final class

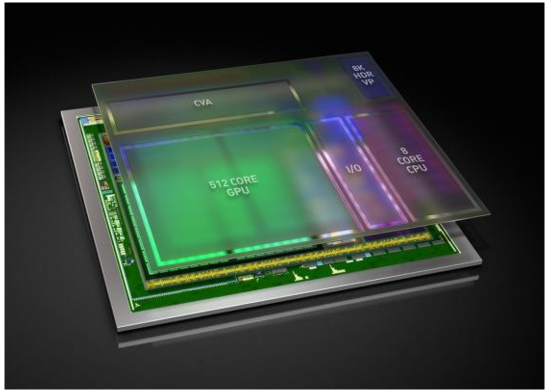
Weighted average safeness is >99 %



SAFETY SUPPORT IN NVIDIA GPUS

SYSTEMATIC DEVELOPMENT OF GPU HARDWARE

Selected GPU cores targeted for automotive usage are developed with a process for ISO 26262 compliance



LAYERED SAFETY MECHANISMS

Redundant execution
HW machine checks
Parity/ECC protection of key structures

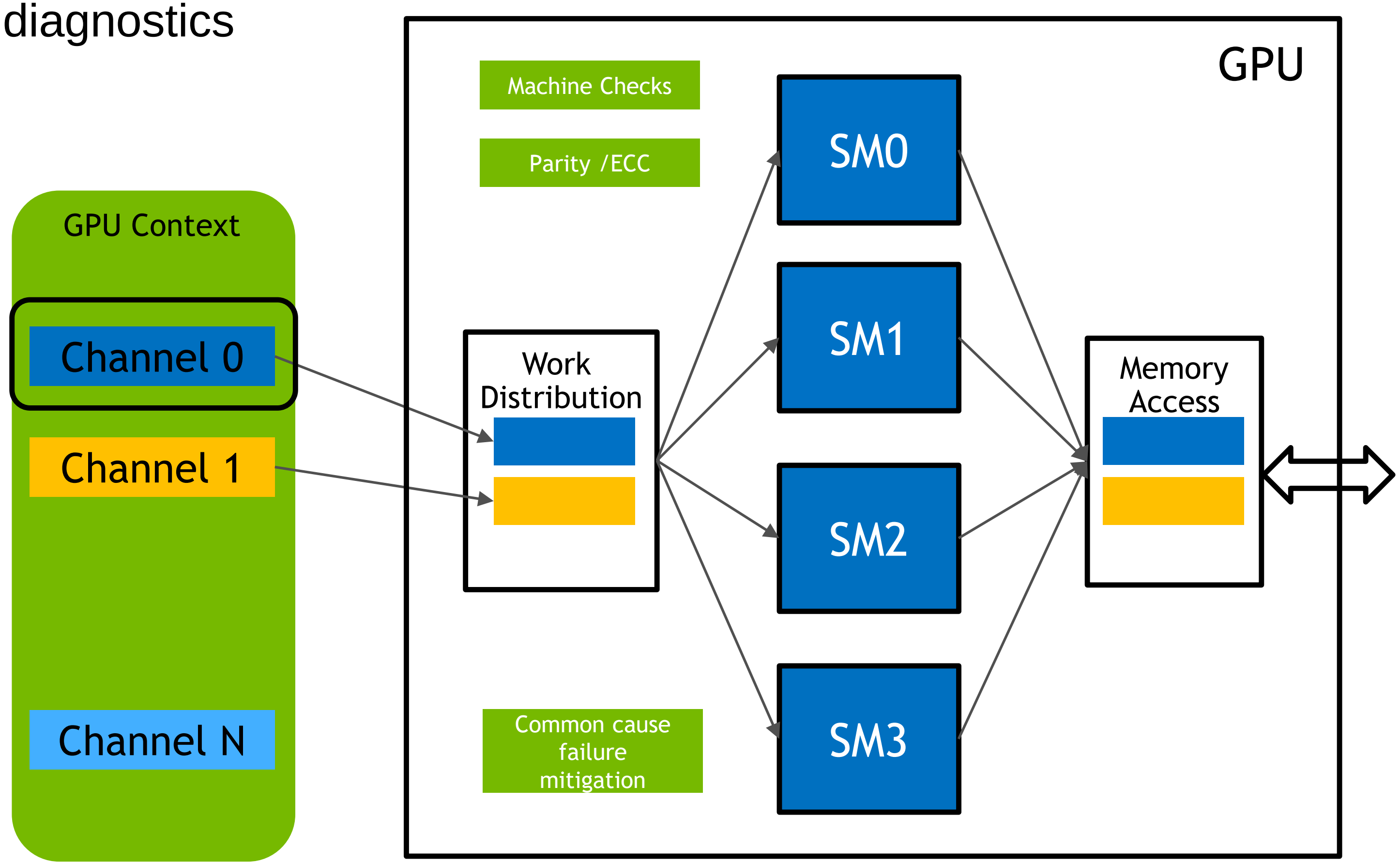
HW plausibility checks enabling multiple execution checks throughout the GPU,

Protection of large safety related memories,

Dependent failure mitigation; mainly caches and shared structures,

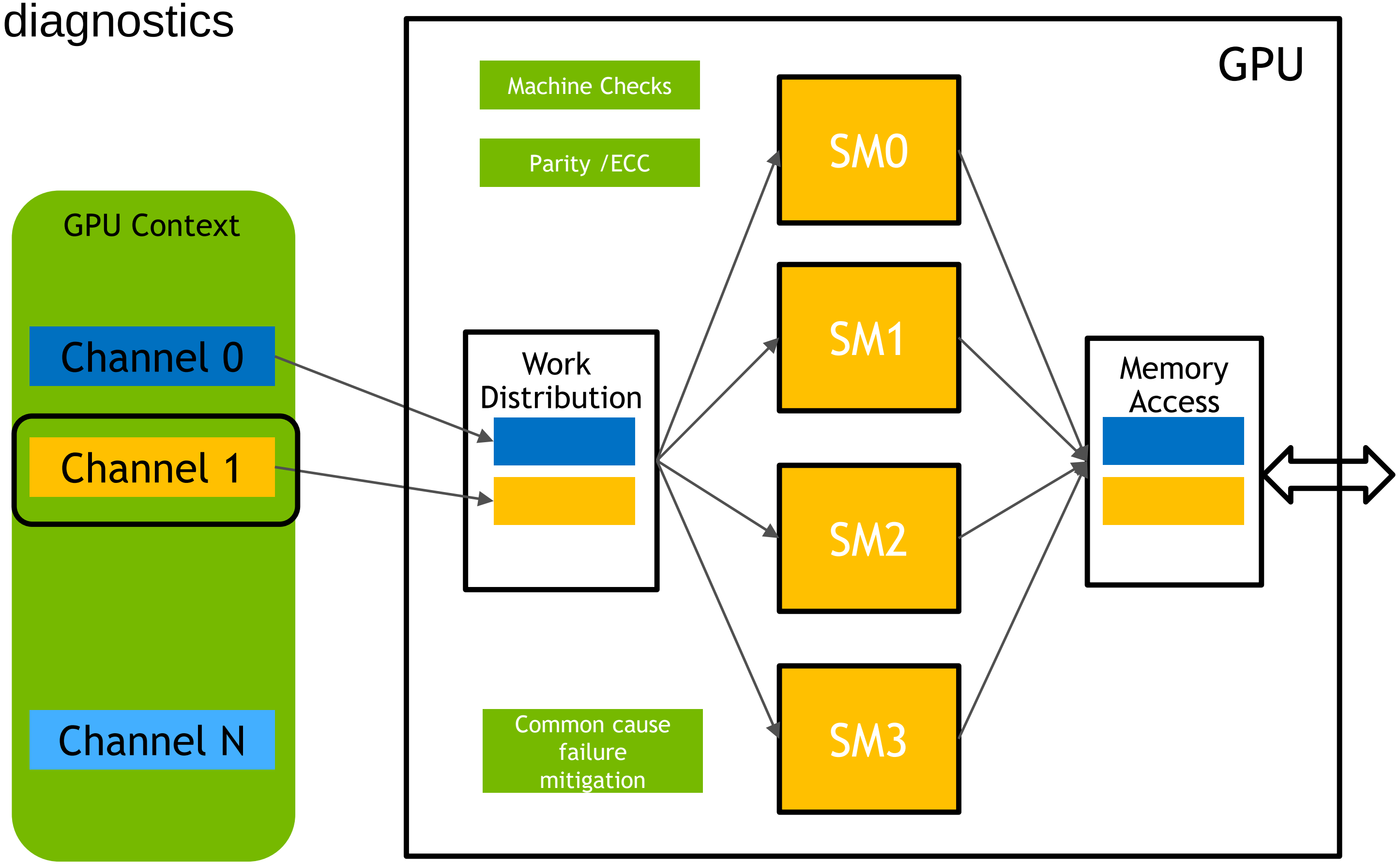
FLEXIBLE REDUNDANCY MODEL

Flexible Execution model
Built-in HW and SW diagnostics



FLEXIBLE REDUNDANCY MODEL

Flexible Execution model
Built-in HW and SW diagnostics



SYSTEMATIC CONSIDERATIONS

Software and tools

Software in the runtime is under development for ISO 26262 compliance

TensorRT

cuDNN

NVIDIA CUDA

Software used in development (training) considered as off-line tools per ISO 26262

GPU FAULT MITIGATION

CONCLUSIONS

Nvidia is developing selected GPUs for compliance to ISO 26262

Nvidia has multiple unique capabilities to analyze safety-related performance of GPUs

Analysis to date indicates DNNs have a high degree of internal redundancy that results in high ratio of safe faults

Selected GPUs are being built with additional hardware and software diagnostic mechanisms

Nvidia is developing software and tools needed to support safety related development

NVIDIA