SAND2014-18451PE

Exceptional service in the national interest

Sandia National Laboratories

Hardware-based Intrusion Detection for Critical Embedded Systems

Nathan J. Edwards, Senior Member of Technical Staff R&D

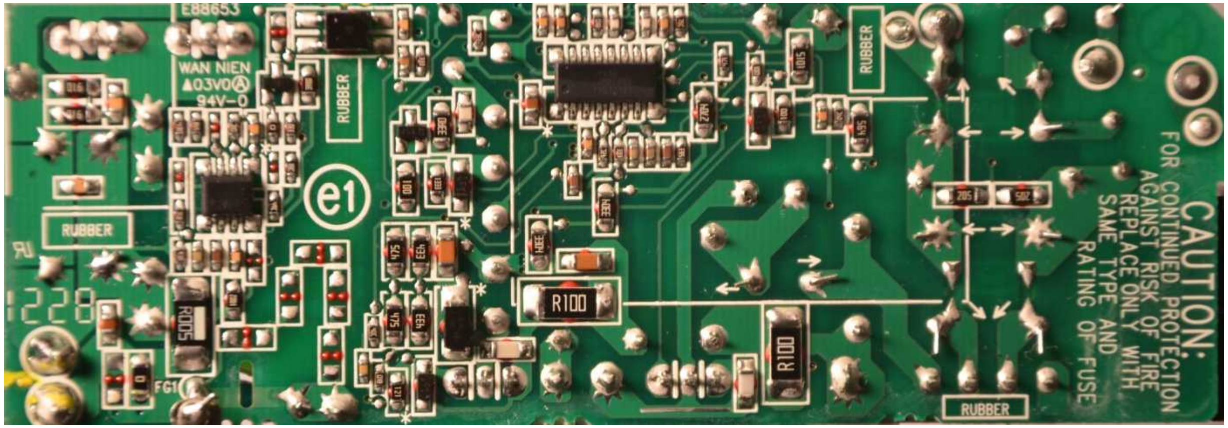
U.S. DEPARTMENT OF ENERGY NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND2014-xxxx

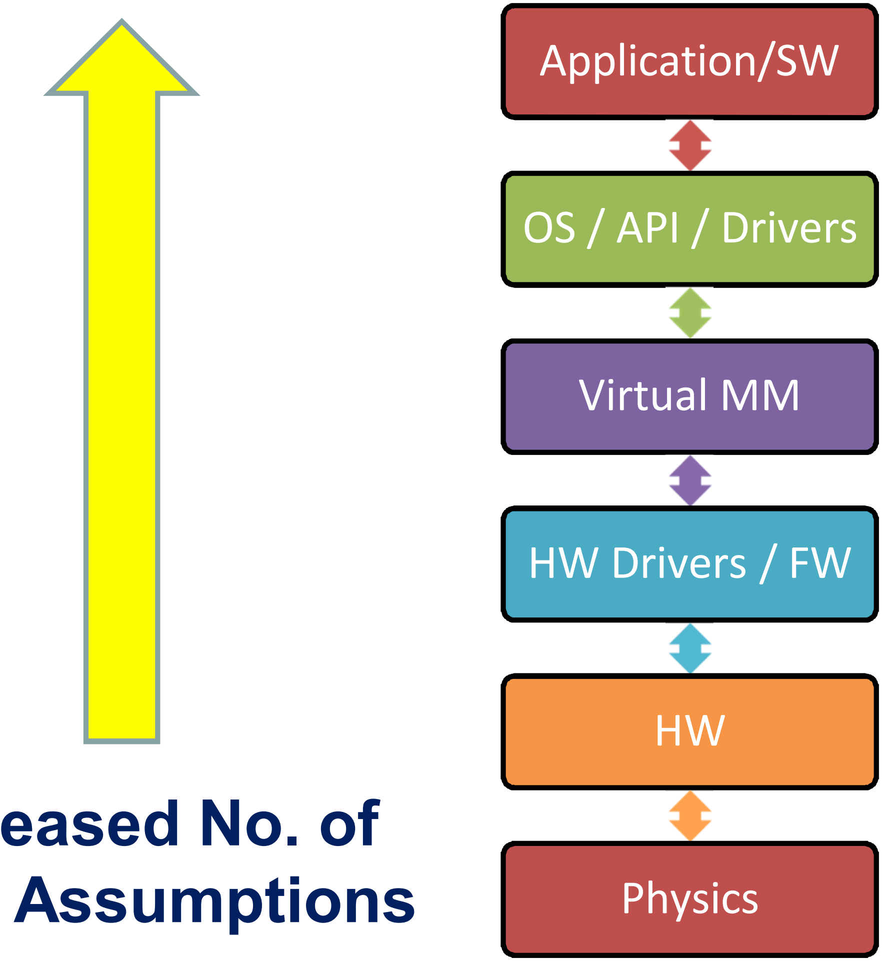
Industrial Control Devices

- PLCs
- PMUs
- Breakers/Relays
- Metering
- RTUs
- Gateways
- Others….

DOE Roadmap to Achieve Energy Delivery Systems Cybersecurity 2011



System Levels of Trust



Application/SW

OS / API / Drivers

Virtual MM

HW Drivers / FW

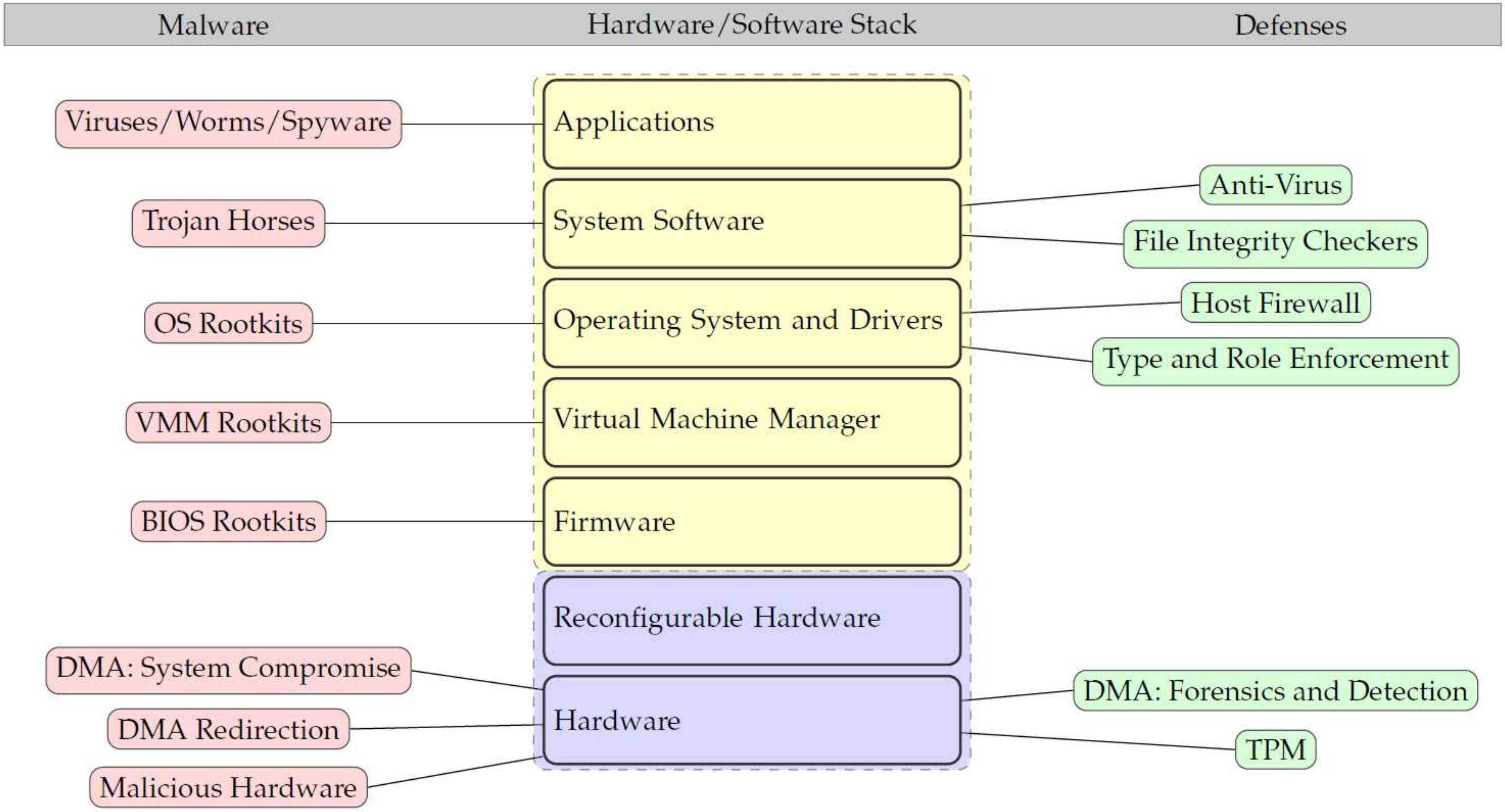
HW

Physics

Increased No. of Trust Assumptions

Data transfer between levels can be risky!

The Current Situation

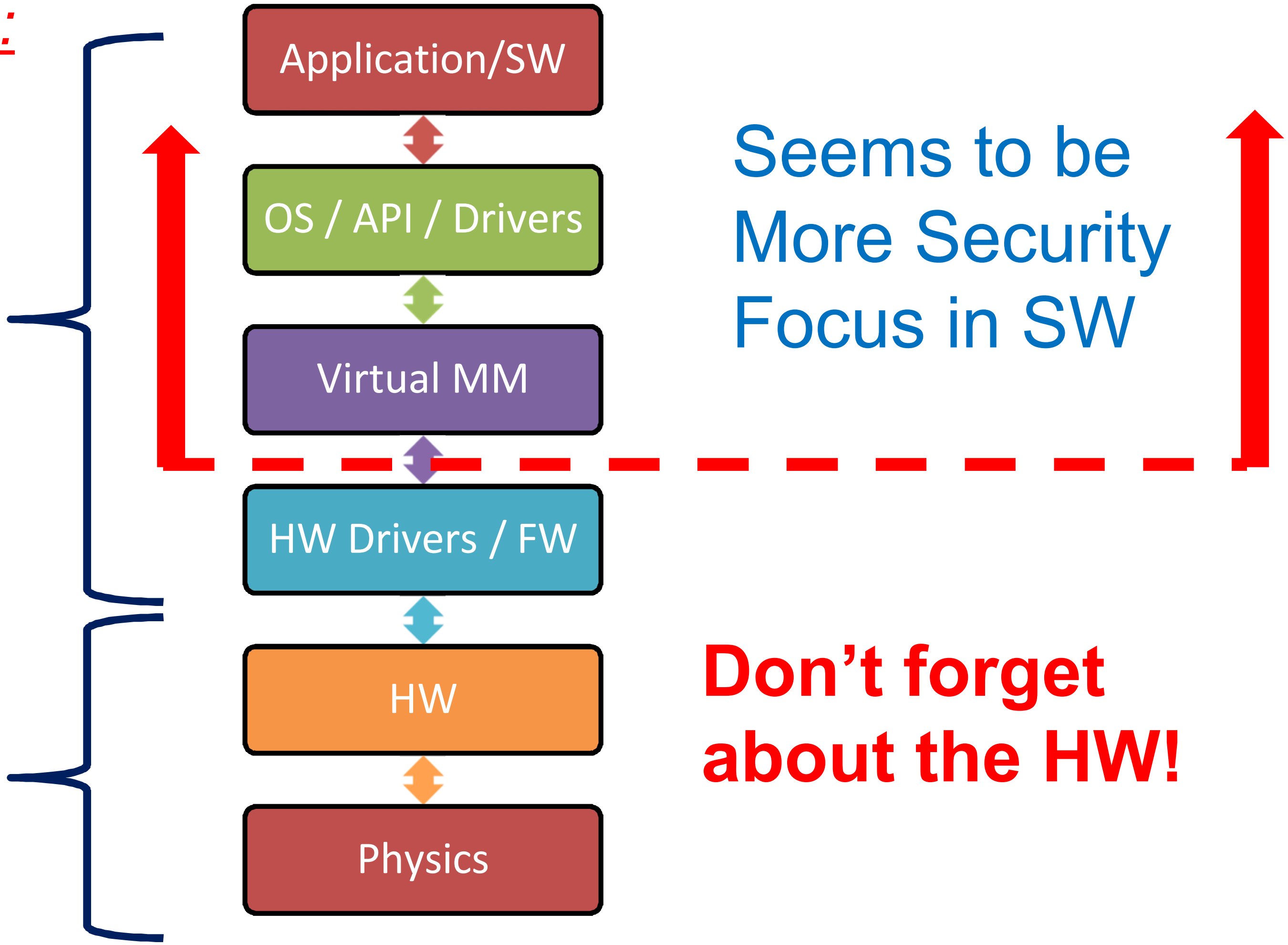


We need more defenses for application logic.

We need more defenses at the lower layers of the hardware/software stack.

System Levels of Trust

Hidden Malware:



Application/SW

OS / API / Drivers

Virtual MM

HW Drivers / FW

HW

Physics

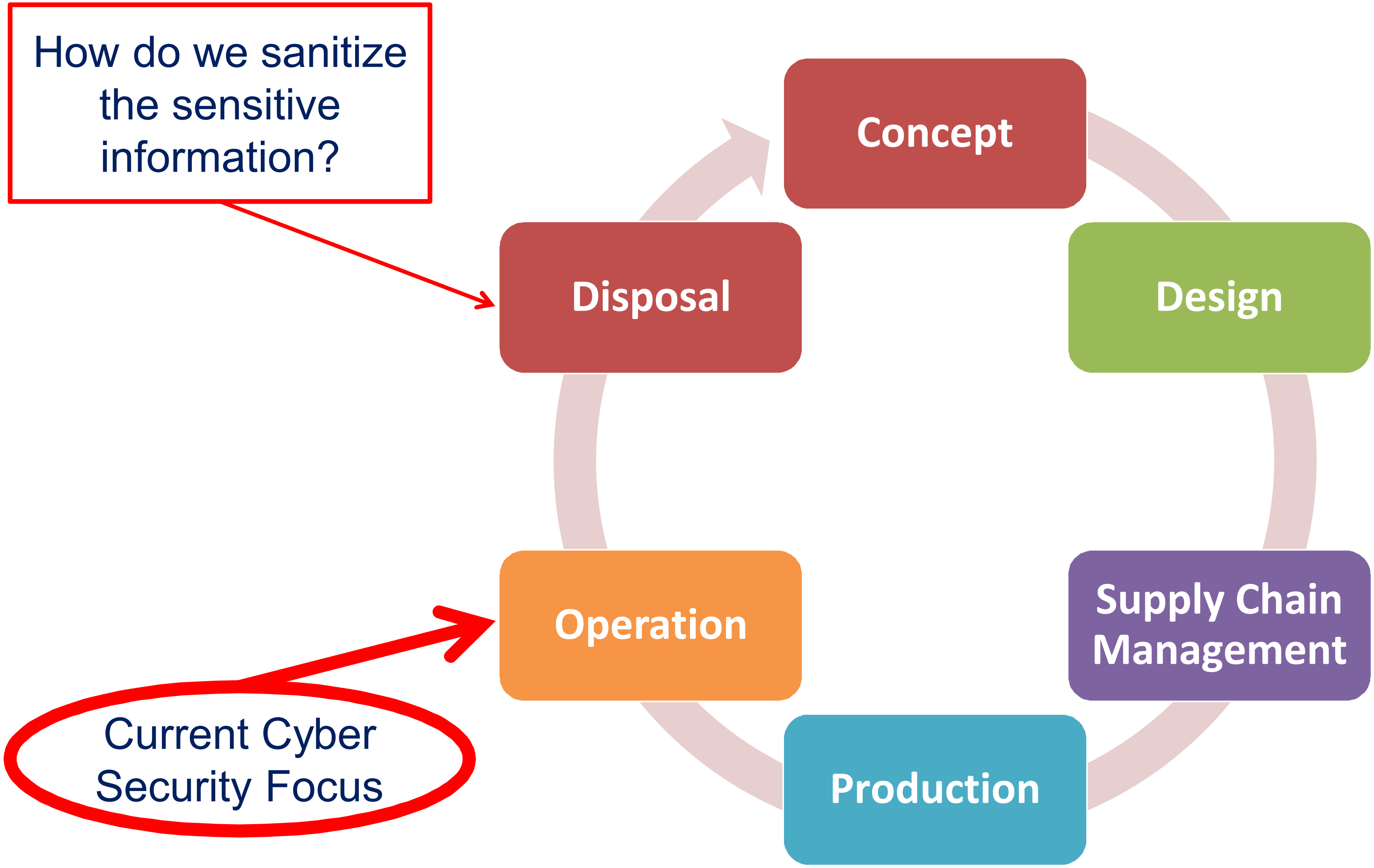
Compiler

Hardware Synthesis

Seems to be More Security Focus in SW

Don’t forget about the HW!

Phases of a Device Lifecycle



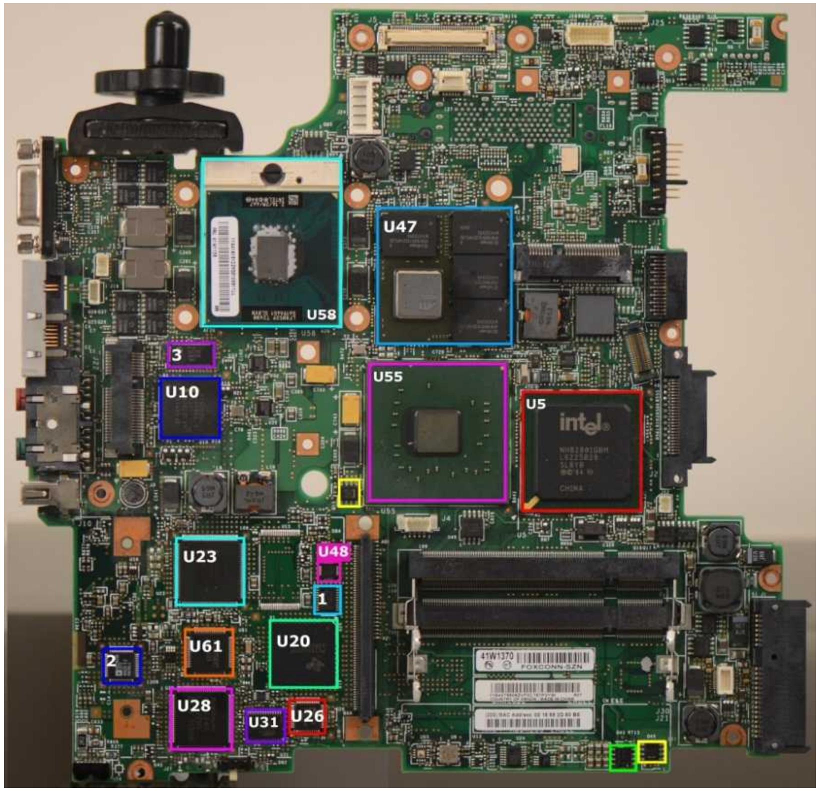
How do we sanitize the sensitive information?

Current Cyber Security Focus

Don’t Forget About the Hardware!

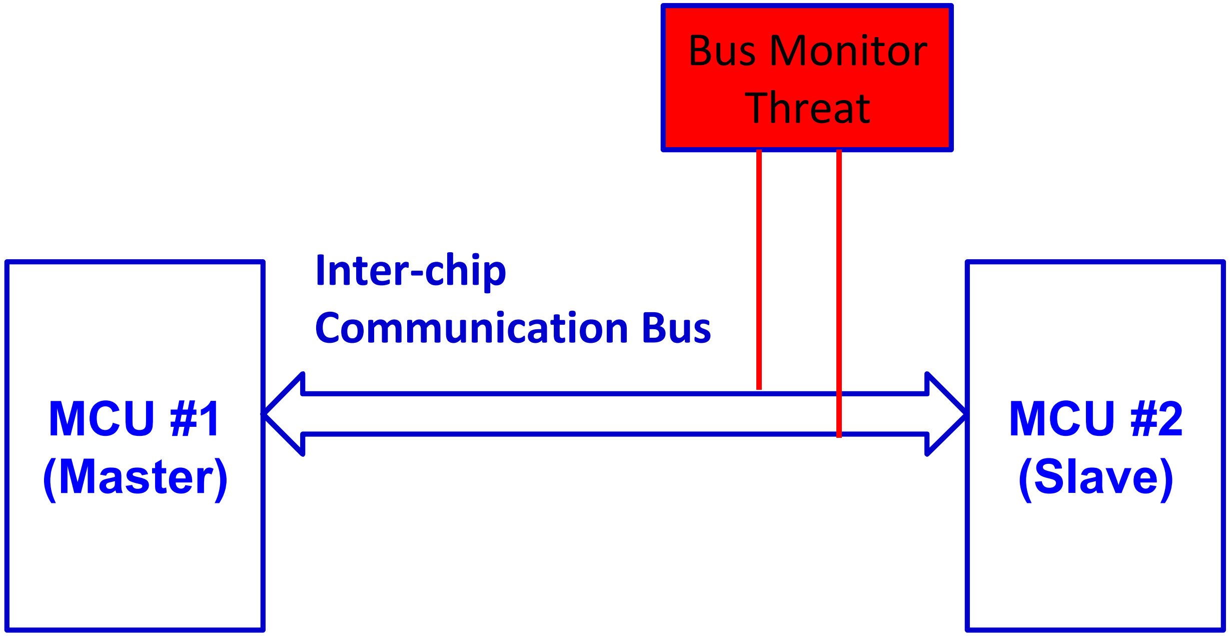
We expect the hardware to execute the instructions we give it.

- How do we know that the results are not copied and sent?
- How do we know that the hardware is not leaking information?
- How do we know that a persistent backdoor has not been inserted?



Hardware Intrusion Model

- ***Passive Attack:*** Inter-chip communication eavesdropping.
- ***Active Attack:*** Communication bus pirating.



Intruder might be able to Acquire Power Usage Data, Activate System, Use Mesh Network, HAN Intrusion

Existing Mechanisms to Detect Hardware Trojans

- **On-chip Trojans:**
  - Very difficult & costly
  - Automated Test Pattern Generation (ATPG)
  - Signal processing using Discrete Hilbert Transforms (DHT and DFHT)
- **Circuit Board Level:**
  - Functional V & V Testing
  - Photographic Identification
  - Side Channel Signal Analysis

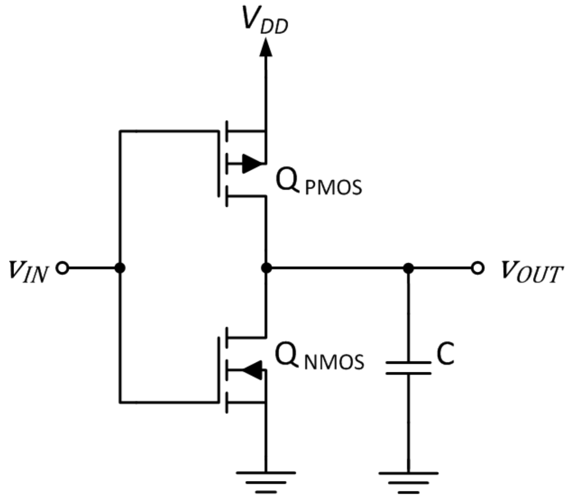
***Most of these are Off-line and not applicable to a fielded/deployed device.**

Hardware Intrusion Research Questions

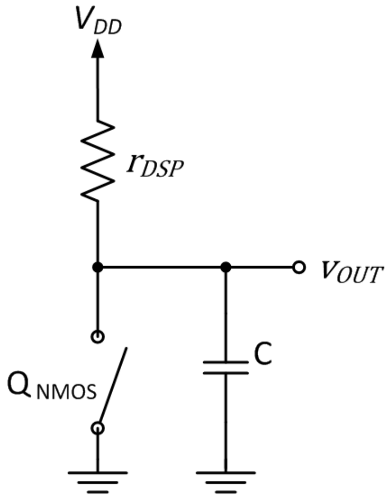
- Can we detect hardware Trojans ?
- Can our detection mechanisms provide any additional information that help characterize the hardware Trojan?
- Can we distinguish between Trojan classes?
- How do we do this?

Hardware Intrusion Model: A Closer Look at Active Attack HW

CMOS Inverter with Capacitor



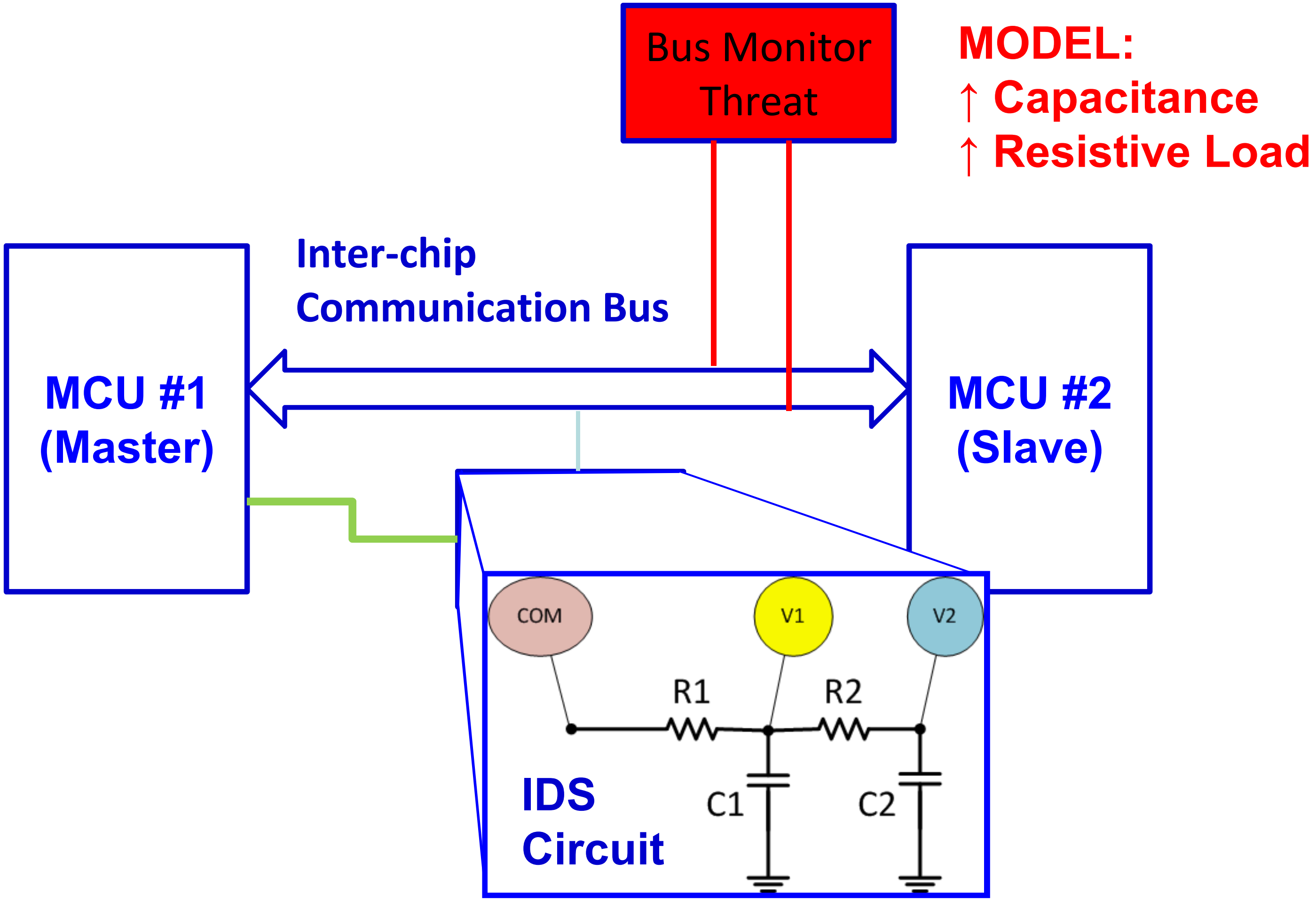
Equivalent Circuit – Logic High



Dynamic Power:

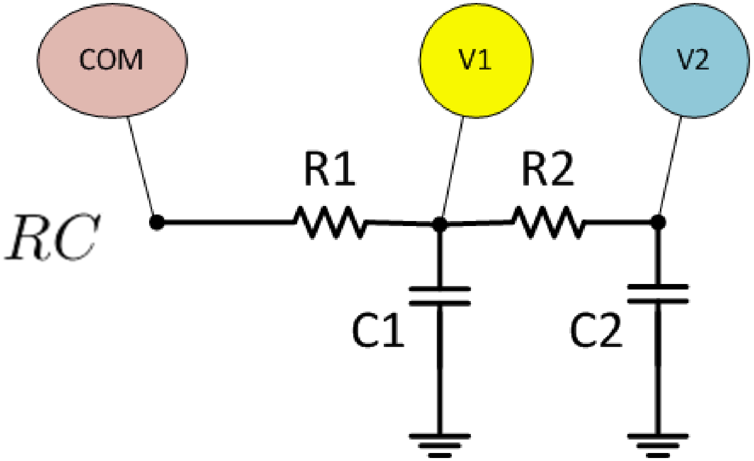
$$\omega_Q = P_D = fCV_{DD}^2, \quad f = \text{transistor switching rate}$$

Hardware Intrusion Detection



IDS Circuit – Two Stage Low-Pass RC Filter

Voltage & Power of IDS Circuit:



$$v(t) = V_0 e^{-\frac{t}{\tau}}, \quad \tau = RC$$

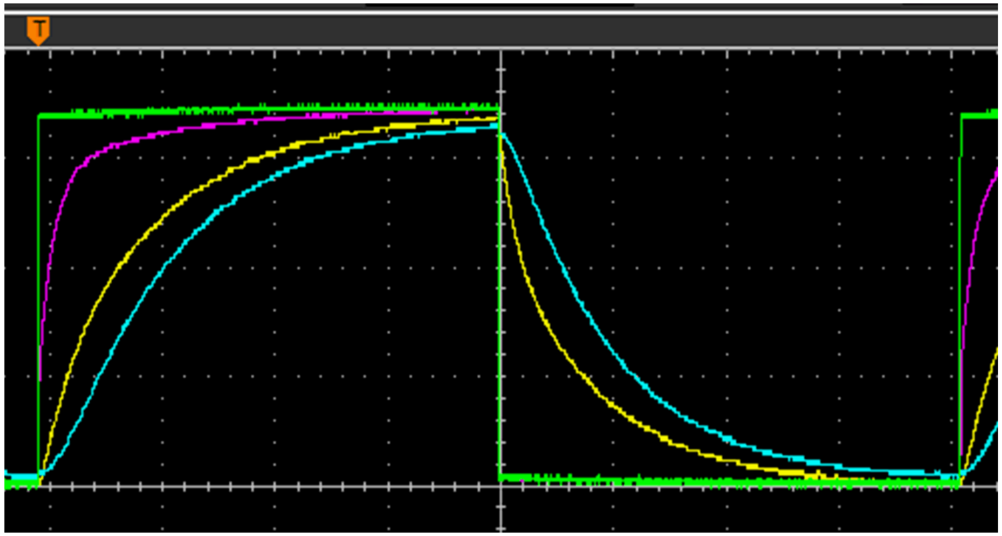
$$\omega_C(t) = \frac{1}{2} C V_t^2$$

$$\omega_R(t) = \frac{1}{2} C V_0^2 (1 - e^{-2t/\tau}), \quad \tau = RC$$

Dynamic Power of Intruder:

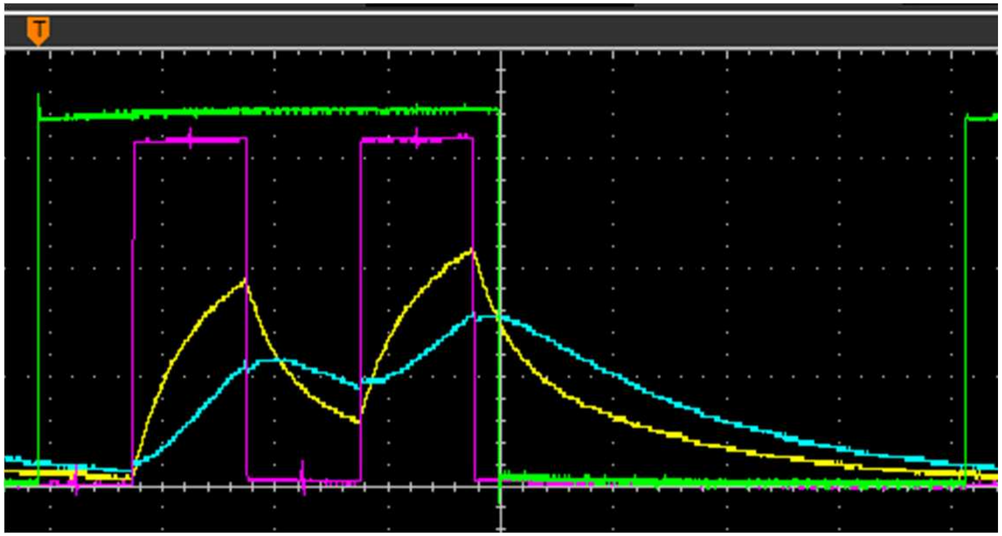
$$\omega_Q = P_D = f C V_{DD}^2, \quad f = \text{transistor switching rate}$$

IDS Voltage Response Signals

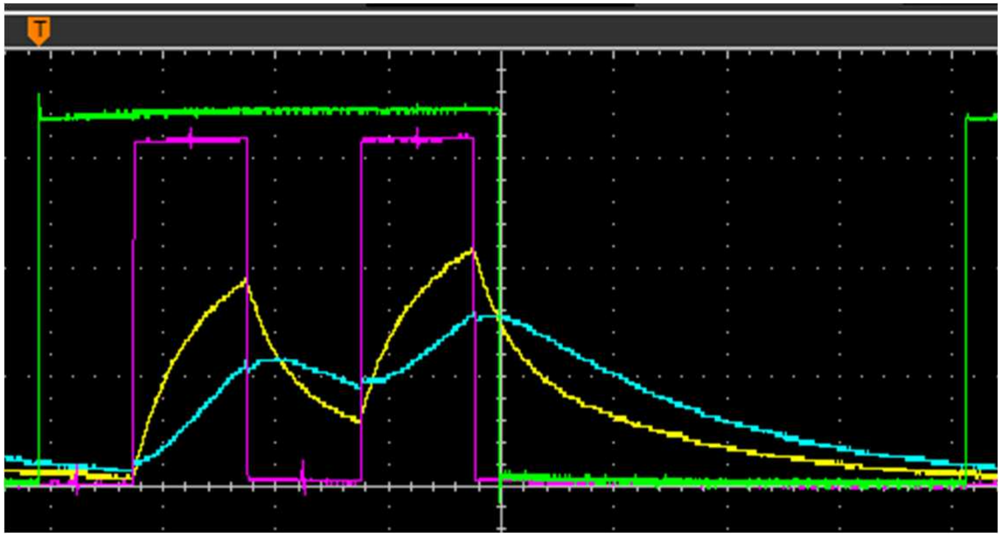


No Intruder

Intruder



Intrusion Detection System Metrics



Voltage Measurements:

- "V1pk"
- "V2pk"
- "V1pkToIDSoff"
- "V2pkToIDSoff"

Interval Slope:

- "SDslopeV1_OnOff“
- "SDslopeV2_OnOff"
- "SlopeV1qty"
- "SlopeV2qty"

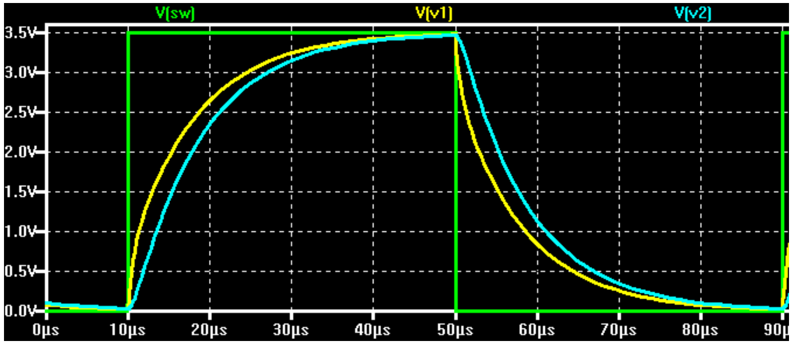
Area Under Curves:

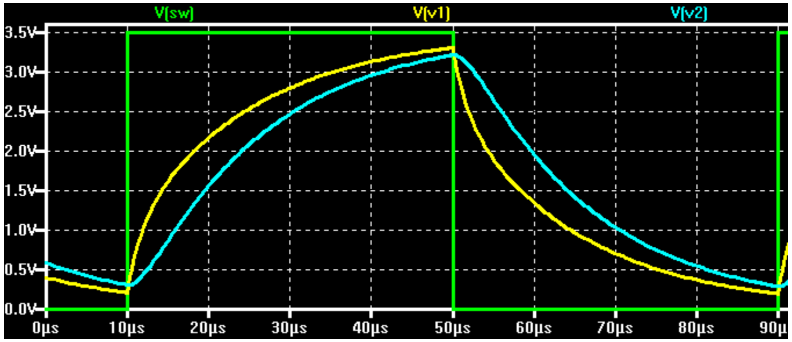
- "AreaV1on"
- "AreaV2on"
- "AreaV1V2on"
- "AreaV1off"
- "AreaV2off"
- "AreaV1V2off"
- "AreaV1V2_OnOff
- "AreaV1_OnOff"
- "AreaV2_OnOff"

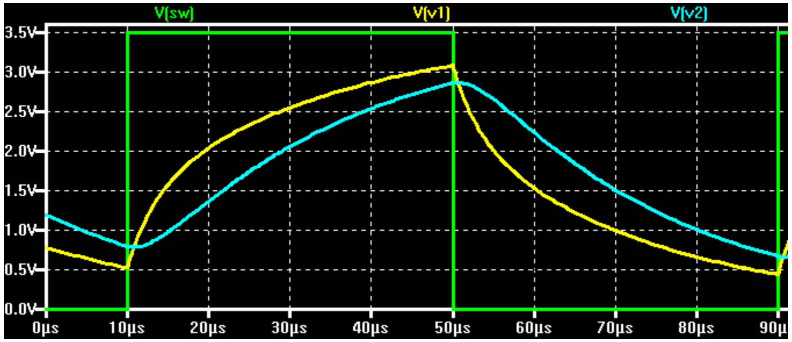
Discrete Components:

- "Cap"
- "Res"

Design of Experiment – IDS Modules







Capacitor Nominal Value

10pF	20pF	39pF	100pF	200pF
84.5KΩ	64.9KΩ	49.9KΩ	21.5KΩ	12.4KΩ
165KΩ	143KΩ	97.6KΩ	49.9KΩ	24.9KΩ
249KΩ	210KΩ	165KΩ	84.5KΩ	45.3KΩ

Resistor Values

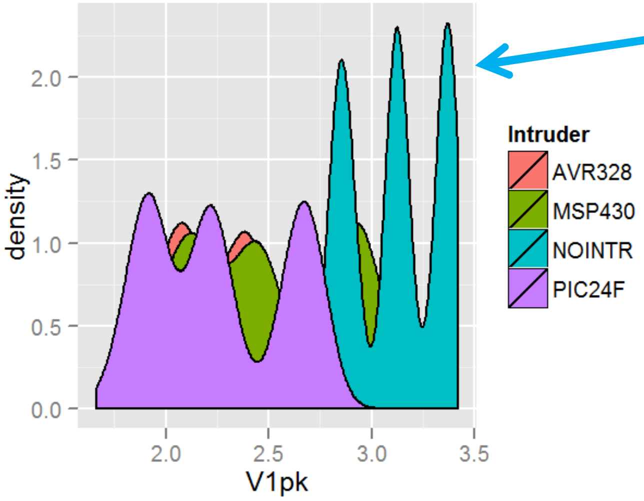
- Charge Cycle = 40us (4bits @100kHz)
- Intruders use SPI or GPIO hardware to attack system

System Noise Characterization

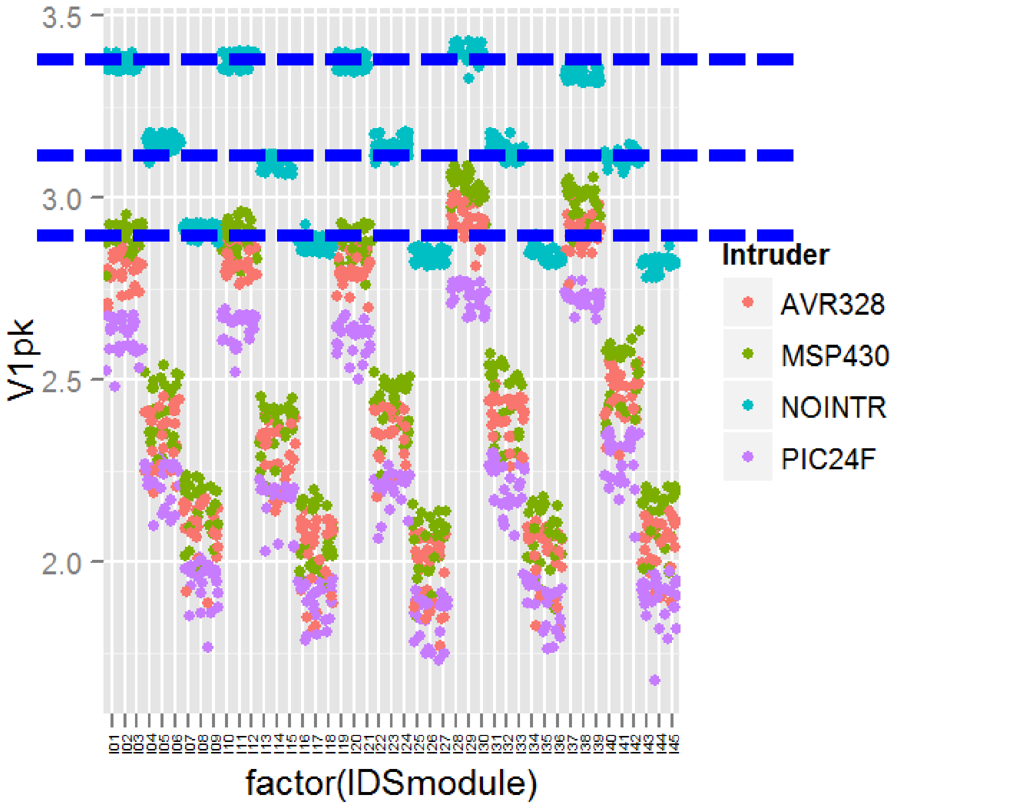
- For threshold-based IDS algorithms, it is critical to understand the level(s) of system noise
- Isolated data for NOINTR
- Primarily looked at average std. deviations since the arithmetic means for each RC combination is different

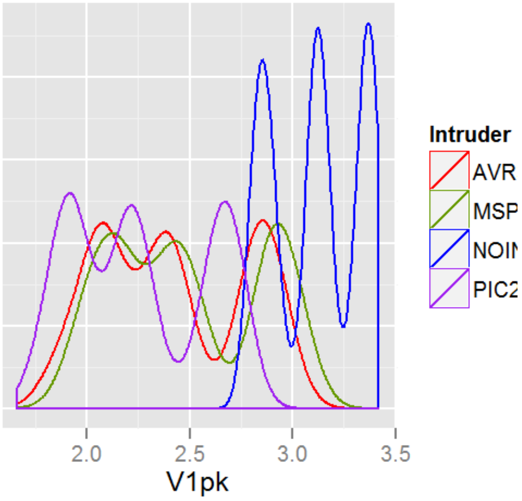
Metric	SD % of Mean	Units
V1pk	0.51%	V
V2pk	0.57%	V
V1pkToIDSoff	769.56%	s
V2pkToIDSoff	146.59%	s
AreaV1on	0.99%	V*s
AreaV2on	1.16%	V*s
AreaV1V2on	1.48%	V*s
AreaV1off	1.32%	V*s
AreaV2off	0.79%	V*s
AreaV1V2off	1.78%	V*s
AreaV1V2_OnOff	372.60%	V*s
AreaV1_OnOff	16.74%	V*s
AreaV2_OnOff	43.25%	V*s
SDslopeV1_OnOff	417.87%	V/10µs
SDslopeV2_OnOff	279.48%	V/10µs
SlopeV1qty	1.63%	integer
SlopeV2qty	7.34%	integer

Graphical Analysis – V1pk

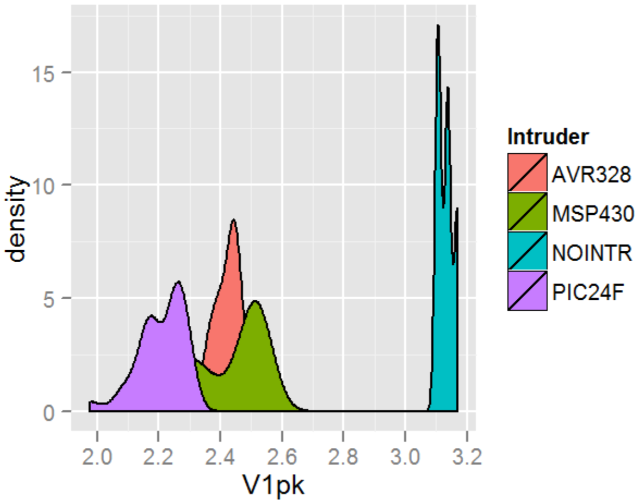


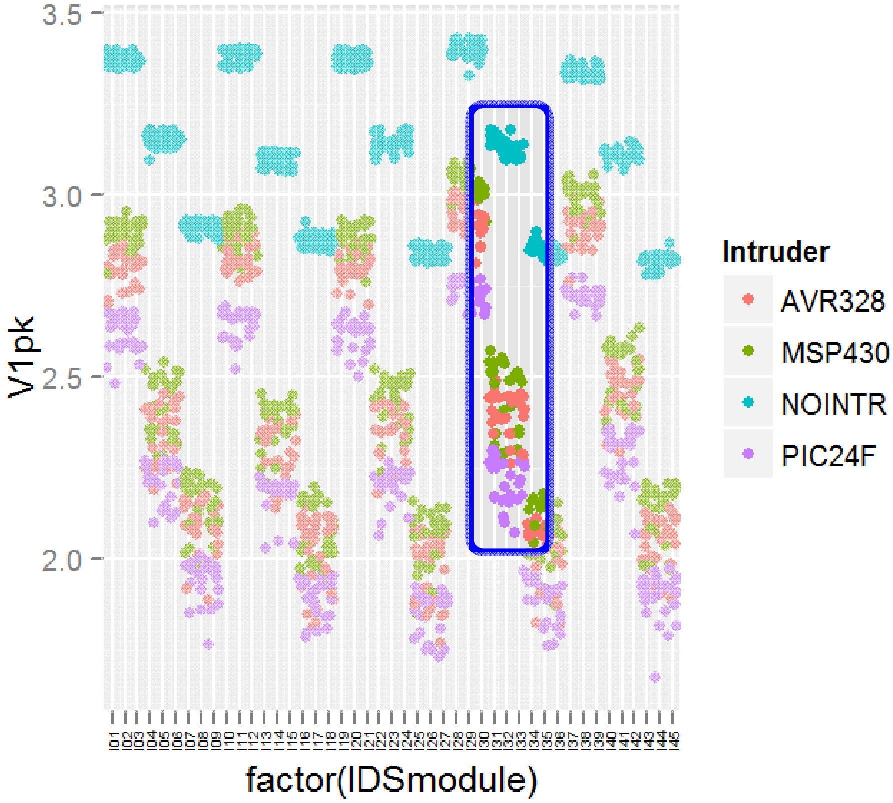
Looks like 3 modes in Density Plot

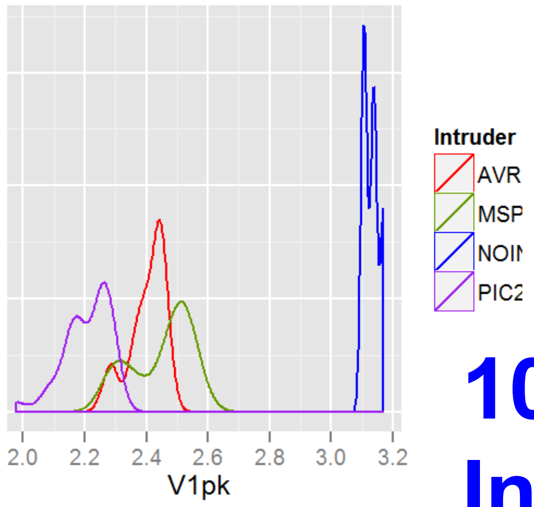




Graphical Analysis – V1pk (single mode)

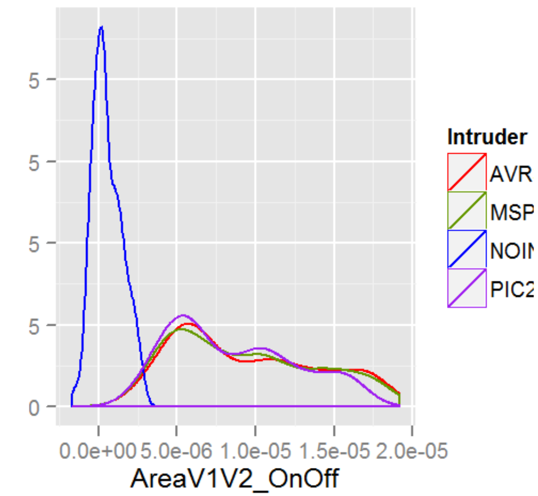


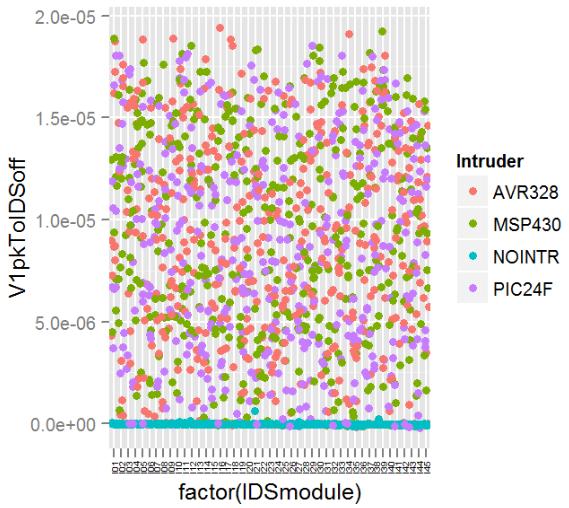


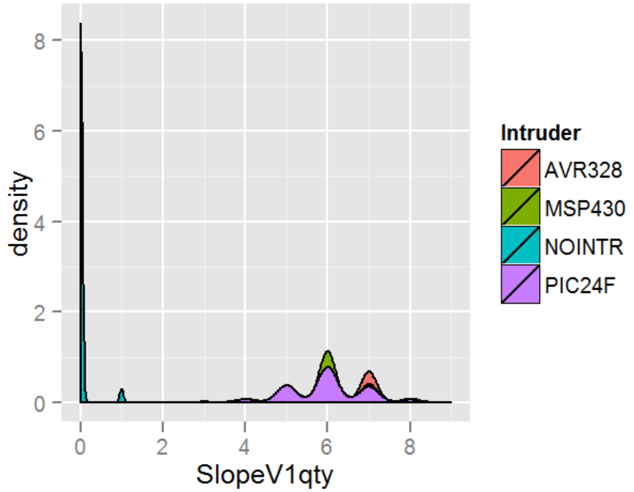


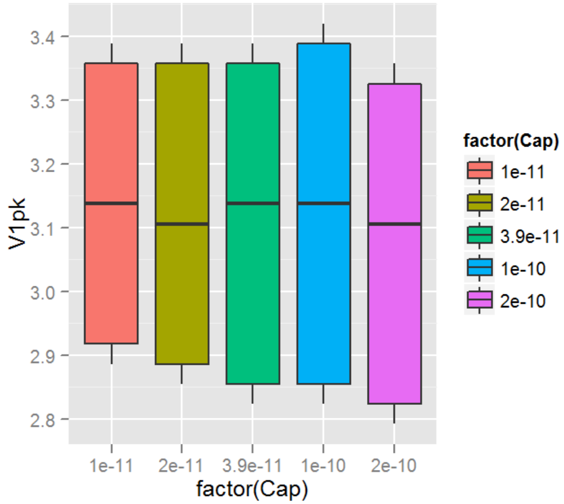
100% Identification of Intruder vs no-intruder

Some Factors Are Not Useful…









IDS Model with Coefficients

Logistic Regression Model:

$$logit(\pi_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{n-1} x_{n-1}, \qquad \beta_0 = \alpha$$

- Helps answer the binary question “Intruder vs No-intruder”
- Logistic Regression can be multinomial.
- Fast processing in embedded systems (similar to linear regression).
- Can apply Bayesian approach to Logistic Regression.

Building and Refining the IDS Model

- Observed metrics become “predictors”.
- Goal is to have a model that is accurate, but not too computationally intensive.
- Use backward elimination and compare several statistics to decide the next predictor to remove.

$$Deviance = D = -2ln\left(\frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}}\right)$$

$$LRstat = G = D(\text{model without the variable}) - D(\text{ model with the variable})$$

$$AIC = 2k - 2ln\,(L) = \chi^2 + 2k$$

Building and Refining the IDS Model

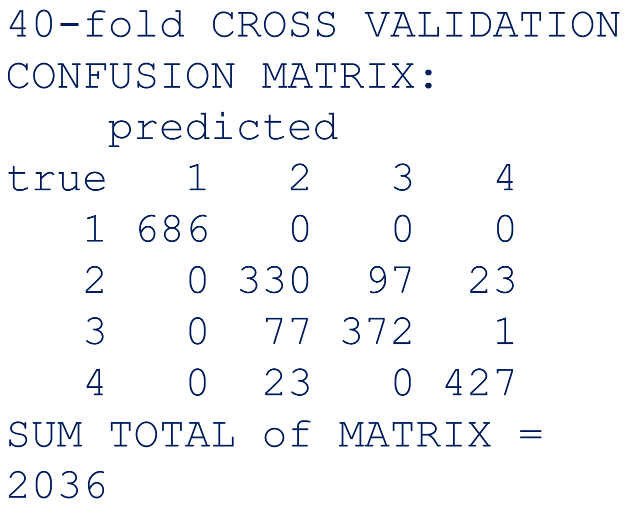
1. Start with full model that includes all 19 predictors.
2. Compare reduced model to full model.
3. ANOVA on Deviance.
4. Select next predictor to remove.

```
AIC: 1025.812
 Resid. df Resid. Dev   Test    Df  LR stat.   Pr(Chi)
1      6054   917.8122
2      6051   917.4638 1 vs 2     3 0.3483742 0.950688
Analysis of Deviance Table (Type II tests)

Response: Intruder
                LR Chisq Df Pr(>Chisq)
IDSmodule          80.10  3  < 2.2e-16 ***
Cap                 9.88  3  0.0195772 *
Res               139.90  3  < 2.2e-16 ***
V1pk              578.27  3  < 2.2e-16 ***
V2pk              196.09  3  < 2.2e-16 ***
V1pkToIDSoff       25.65  3  1.129e-05 ***
V2pkToIDSoff        9.37  3  0.0247645 *
AreaV1on          132.26  3  < 2.2e-16 ***
AreaV2on          145.05  3  < 2.2e-16 ***
AreaV1off          24.56  3  1.911e-05 ***
AreaV2off          48.25  3  1.888e-10 ***
AreaV1V2_OnOff     51.51  3  3.815e-11 ***
SDslopeV1_OnOff     4.63  3  0.2010404
SDslopeV2_OnOff    56.38  3  3.490e-12 ***
SlopeV1qty         29.36  3  1.883e-06 ***
SlopeV2qty         16.96  3  0.0007219 ***
AreaV2_OnOff       30.66  3  1.002e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
```

IDS Model Performance Metrics

- 40-fold cross validation
- Confusion Matrix & Overall Accuracy
- TPR, FPR, Precision, K-hat



```
40-fold CROSS VALIDATION
CONFUSION MATRIX:
     predicted
true    1    2    3    4
   1  686    0    0    0
   2    0  330   97   23
   3    0   77  372    1
   4    0   23    0  427
SUM TOTAL of MATRIX =
2036
```

$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$False\ Positive\ Rate = \frac{FP}{FP + TN}$$

$$Precision = PPV = \frac{TP}{TP + FP}$$

$$\hat{K} = \frac{p_O - p_C}{1 - p_C} = \frac{actual\ agreement - chance\ agreement}{1 - chance\ agreement}$$

IDS Model: Goodness of Fit & Performance

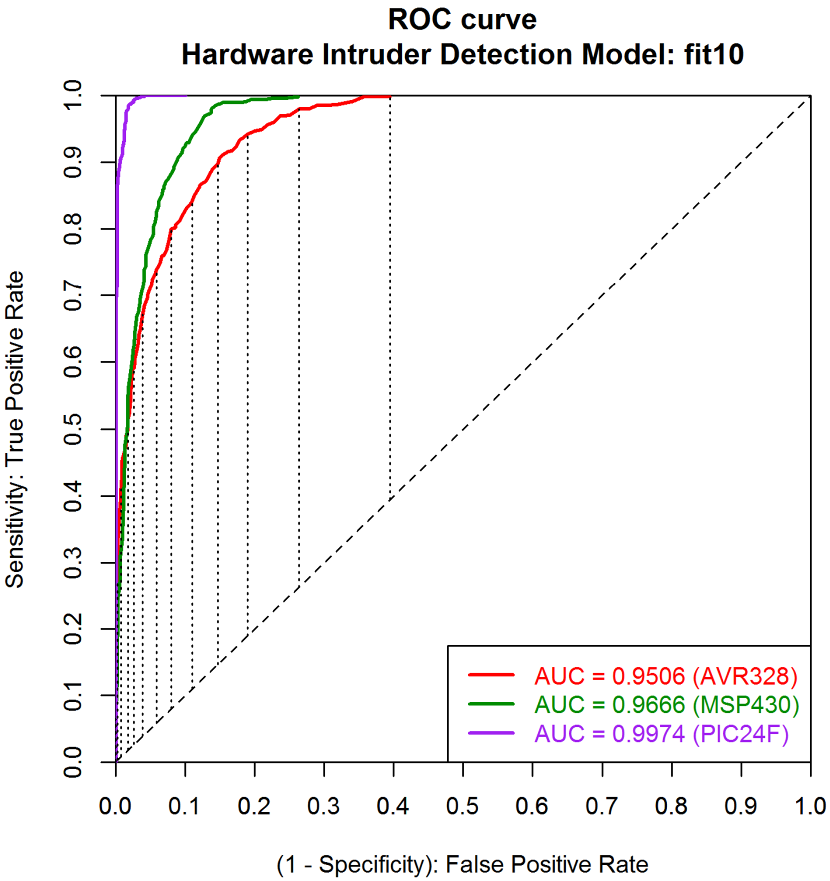
Model	Dev G^2	df	Pr(Chi)	AIC	Overall Accuracy	Mean TPR	Mean FPR	Mean PPV	$\hat{K}$ Statistic
fitF	4650.7	54	0	1031.46	0.8939	0.8800	0.0441	0.8794	0.8566
fit1	-5E-04	0	1	1031.46	0.8934	0.8794	0.0443	0.8789	0.8560
fit2	-0.006	0	1	1031.47	0.8939	0.8800	0.0441	0.8794	0.8566
fit3	0.3484	3	0.9507	1025.81	0.8939	0.8800	0.0441	0.8796	0.8566
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
fit9	46.847	18	0.0002	1042.31	0.8919	0.8778	0.0449	0.8768	0.8540
fit10	66.195	21	1E-06	1055.66	0.8915	0.8772	0.0451	0.8769	0.8533
fit11	89.806	24	2E-09	1073.27	0.8861	0.8711	0.0471	0.8702	0.8460
fit12	115.04	27	8E-13	1092.51	0.8875	0.8728	0.0465	0.8721	0.8480
fit13	186.61	30	0	1158.07	0.8767	0.8606	0.0506	0.8601	0.8334
fit14	262.45	33	0	1227.92	0.8654	0.8478	0.0550	0.8469	0.8181

fit10: 89.15% accurate, ↑TPR, ↓FPR, $\hat{K}$ close to Full model

IDS Model Performance:

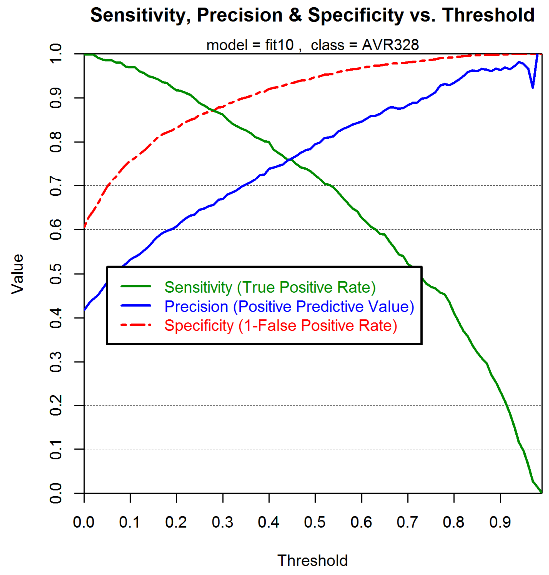
Receiver Operating Characteristic Curve

- Area Under Curve (AUC) provides comparison of model to that of a random guess
- Data set stratification (or lack-of) can be apparent in a ROC curve.

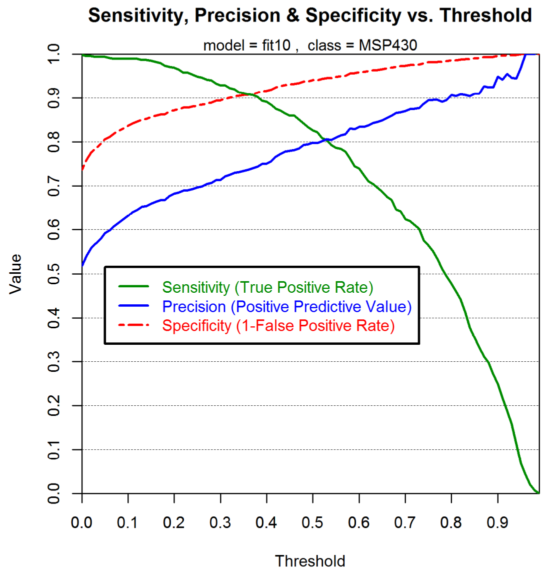


IDS Model Performance:

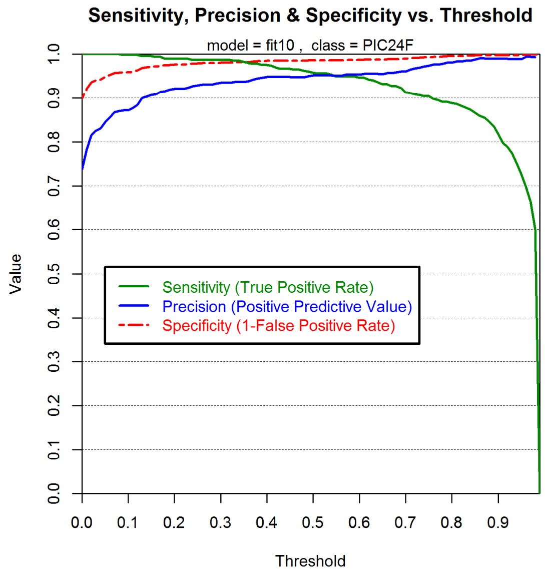
Sensitivity, Precision & Specificity



AVR328



MSP430



PIC24F

Hardware IDS Takeaways

- 100% Identification of Intruder
- Classifier System:
  - 89.15% accurate, 87.7%TPR, 4.5%FPR, $\hat{K}$ = 0.853

- Not a stand-alone solution for all security issues.
- Very cost-effective solution for new capability.
- Can be combined with Specification-based IDS and System-wide IDS for high-resolution and complete security view.
- Starts to address supply-chain hardware security issues.
- Signatures of various intruders are distinct.

Thank You!

The technology and methodology contained herein are subject to one or more pending patents: USPTO 13/834,673; 13/782,808; and 14/494,306.