CS 578: CYBER-SECURITY

PART III: ROWHAMMER

Sanghyun Hong
sanghyun.hong@oregonstate.edu

SAIL
Secure AI Systems Lab

ANNOUNCEMENT

- Do not cheat – will be handled by the university

ANNOUNCEMENT

- HW3 was out
- Checkpoint II presentations will be on 5/21
  - 8-10 min presentation + 3 min Q&A
  - Presentation MUST cover:
    - 1 slide on your research topic
    - 1 slides on your research goal(s)
    - 1-2 slides on your hypothesis and evaluation design
    - 1-2 slides on your preliminary results [very important]
    - 1 slide on your next steps until the final presentation

HOW CAN WE BREAK THE ISOLATION?

WHAT CAN WE DO WITH THE ROWHAMMER VULNERABILITY?

1990: Optimal Brain Damage[1] – Graceful Degradations

: we can remove 60% of model parameters, without the accuracy drop

[1]LeCun et al., *Optimal Brain Damage*, NIPs'90

GRACEFUL DEGRADATION

- **Techniques** that rely on the *graceful degradation*
  - **Pruning**[1] : to reduce the inference cost
  - **Quantization**[2] : to compress the network size
  - **Adding noise**[3] : to improve the robustness against adv. examples

[1]Li et al., *Pruning Filters for Efficient ConvNets*, ICLR'17
[2]Wang et al., *Training Deep Neural Networks with 8-bit Floating Point Numbers*, NeuralPS'18
[3]Zhou et al., *Breaking Transferability of Adversarial Samples with Randomness*, ArXiv'18

GRACEFUL DEGRADATION

- **Techniques** that rely on the *graceful degradation*
  - **Pruning**[1] : to reduce the inference cost
  - **Quantization**[2] : to compress the network size
  - **Adding noise**[3] : to improve the robustness against adv. examples

- **Prior work** showed it is difficult to *cause the accuracy drop*
  - **Indiscriminate poisoning**[4]: blend poisons ≈ **11% drop** (avg.)
  - **Storage media errors**[5] : a lot of random bit errors ≈ 5% drop (avg.)
  - **Hardware fault attacks**[6,7] : a lot of random faults ≈ 7% drops (avg.)

[4]Steinhardt et al., *Certified Defenses for Data Poisoning Attacks*, NeuralPS'17
[5]Qin et al., *Robustness of Neural Networks against Storage Media Errors*, Arxiv'17
[6]Li et al., *Understanding Error Propagation in Deep Learning Neural Network (DNN) Accelerators and Applications*, SC'17
[7]Breier et al., *DeepLaser: Practical Fault Attack on Deep Neural Networks*, Arxiv'18

GRACEFUL DEGRADATION – FALSE SENSE OF SECURITY?

- **Techniques** that rely on the *graceful degradation*
  - **Pruning**[1] : to reduce the inference cost
  - **Quantization**[2] : to compress the network size
  - **Adding noise**[3] : to improve the robustness against adv. examples

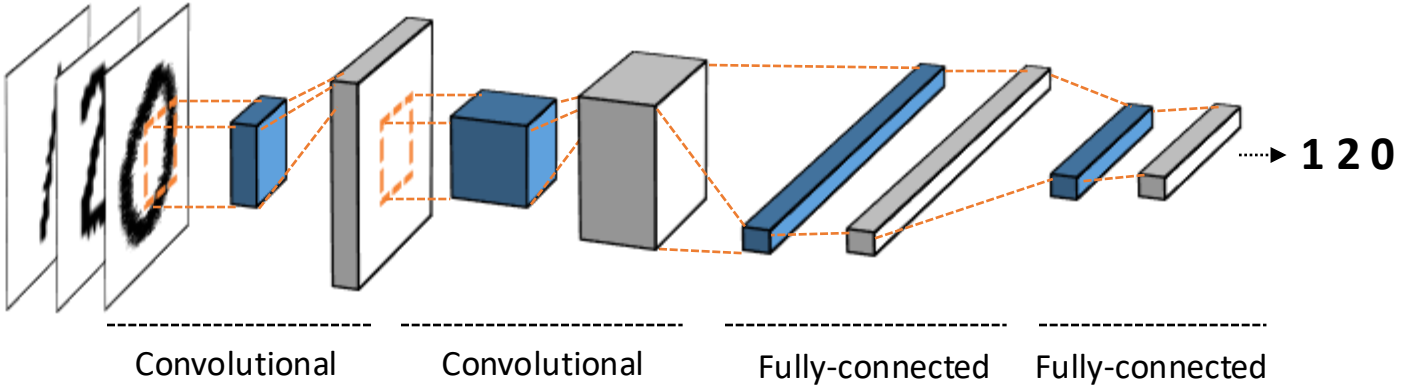
- **Prior work** showed it is difficult to *cause the accuracy drop*
  - **Indiscriminate poisoning**[4]: blend poisons ≈ **11% drop** (avg.)
  - **Storage media errors**[5] : a lot of random bit errors ≈ 5% drop (avg.)
  - **Hardware fault attacks**[6,7] : a lot of random faults ≈ 7% drops (avg.)

They focus on the best-case or the average-case degradation

[4]Steinhardt et al., *Certified Defenses for Data Poisoning Attacks*, NeuralPS'17
[5]Qin et al., *Robustness of Neural Networks against Storage Media Errors*, Arxiv'17
[6]Li et al., *Understanding Error Propagation in Deep Learning Neural Network (DNN) Accelerators and Applications*, SC'17
[7]Breier et al., *DeepLaser: Practical Fault Attack on Deep Neural Networks*, Arxiv'18

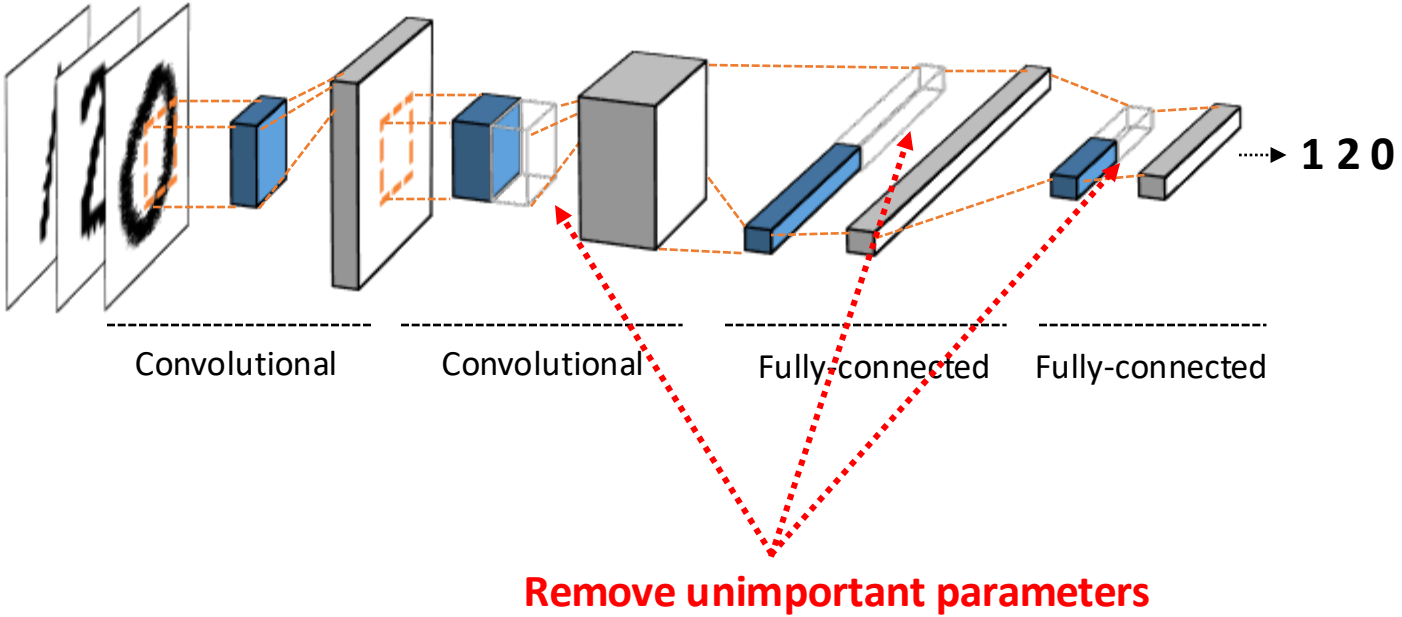
ILLUSTRATION: HOW DNN COMPUTES

- Accuracy: 99%



THE BEST-CASE: OPTIMAL BRAIN DAMAGE[1]

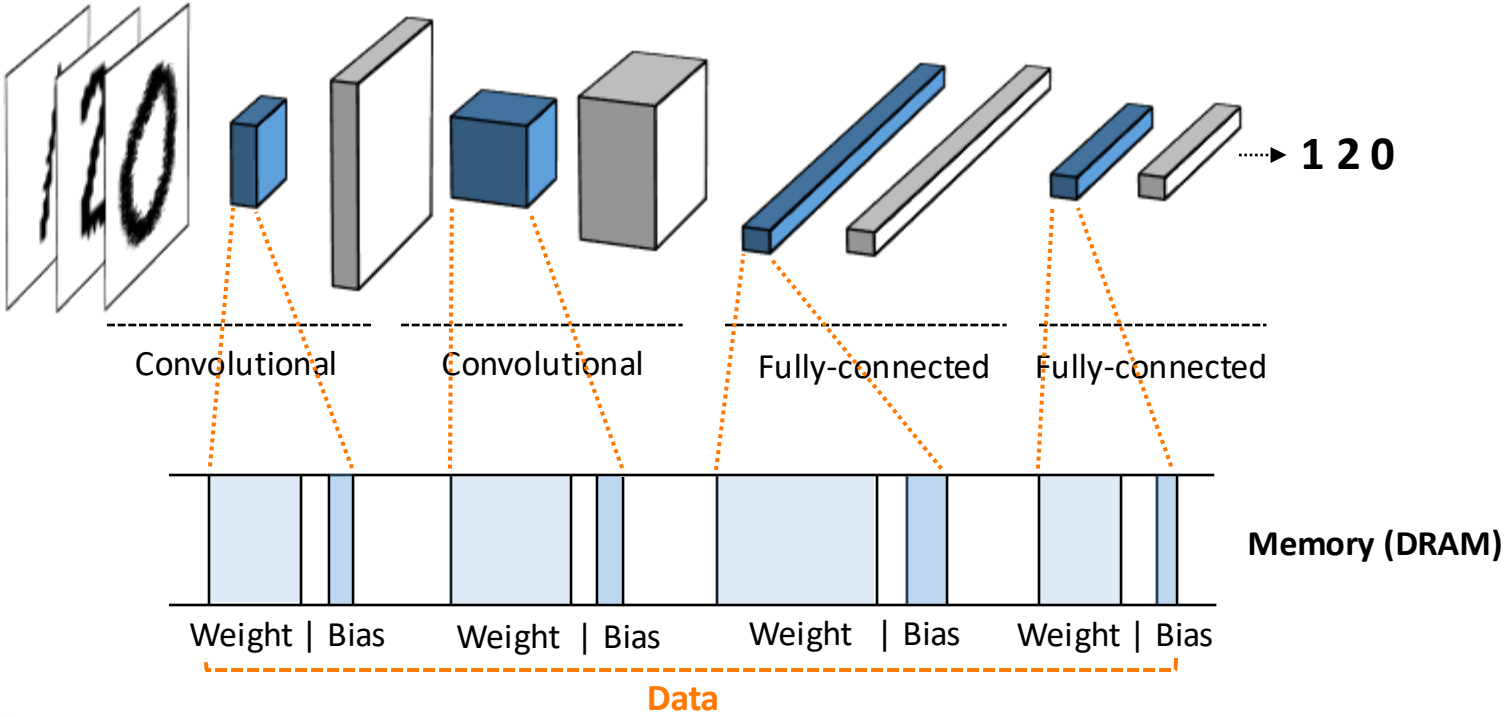
- Accuracy: **99%** (0% drop)



[1]LeCun et al., *Optimal Brain Damage*, NIPs'90

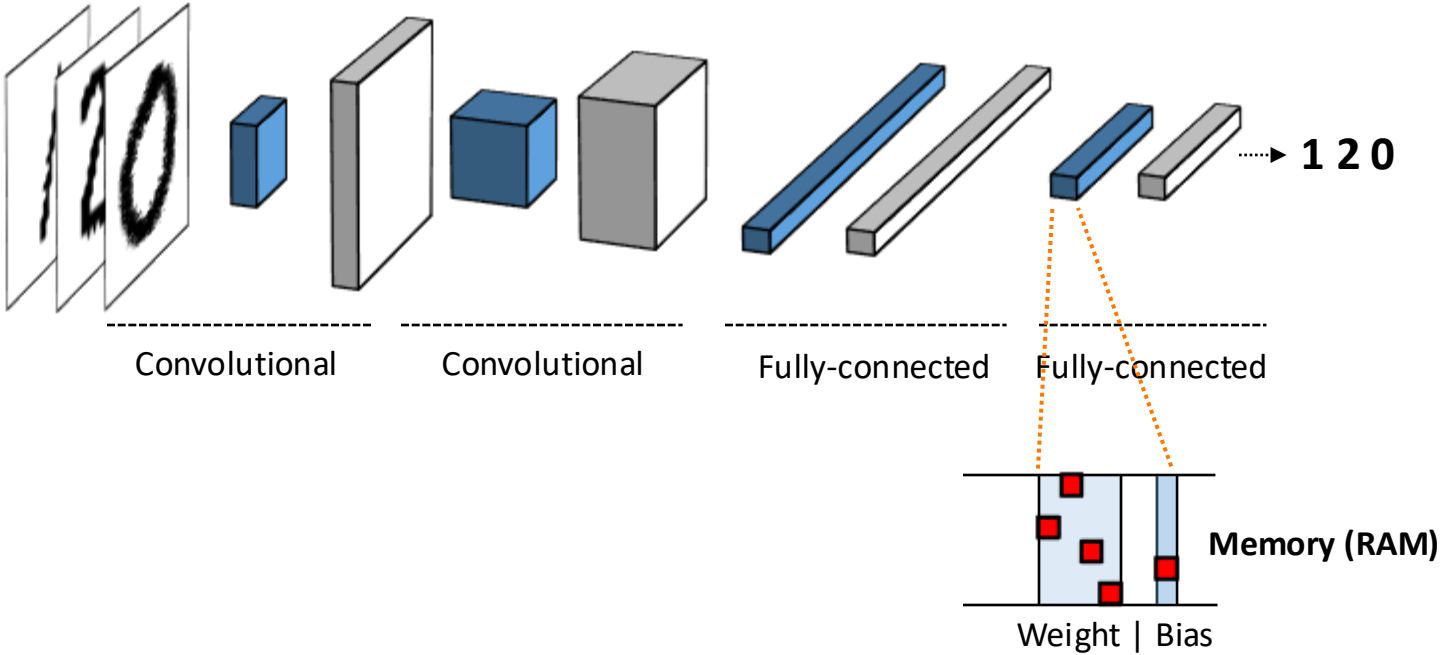
ILLUSTRATION: DNN'S IN-MEMORY REPRESENTATION

- Accuracy: 99%



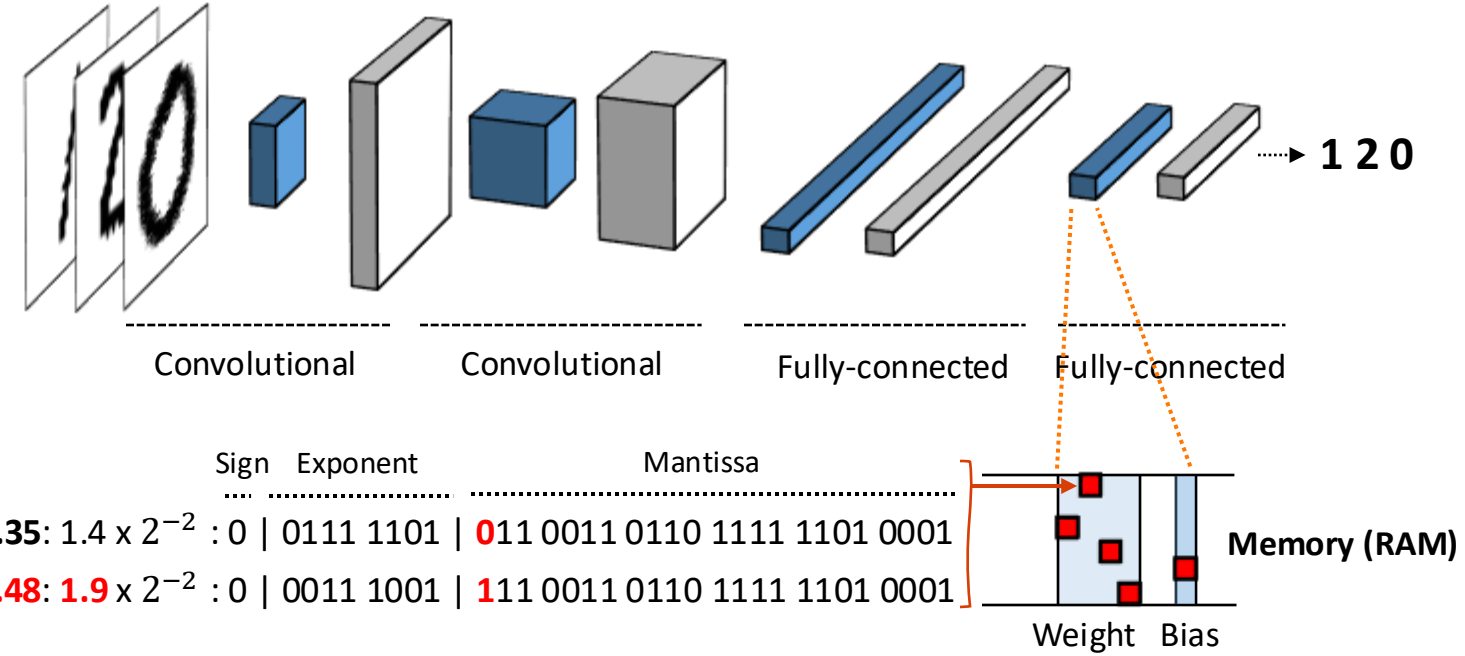
THE AVG-CASE: BITWISE ERRORS IN DNN'S IN-MEMORY REPR.

- Accuracy: **94%** (5% drop on avg.)



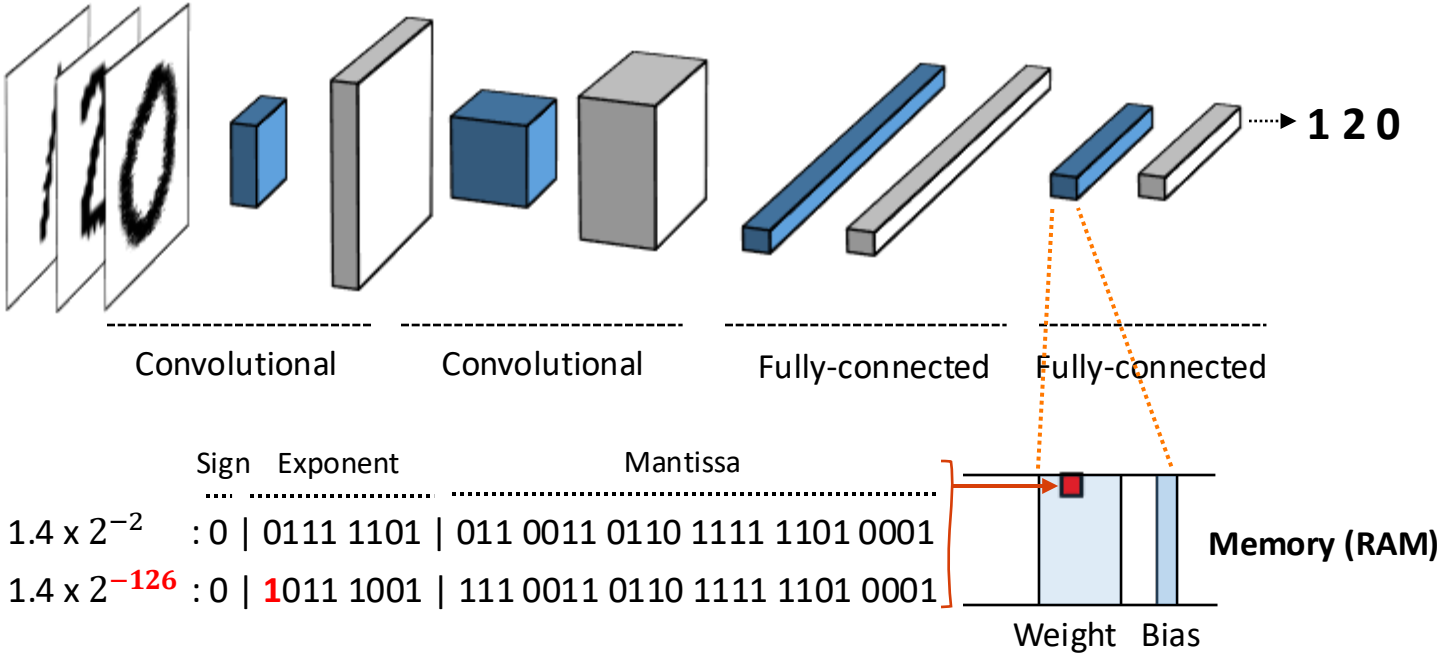
THE AVG-CASE: BITWISE ERRORS IN DNN'S IN-MEMORY REPR.

- Accuracy: **94%** (5% drop on avg.)



THE WORST-CASE: A SINGLE BIT-FLIP

- Accuracy: **58%** (41% drop)



How Vulnerable Are DNNs to a Single Bit-flip?

- **Methodology**
  1) Flip each bit in all parameters of a DNN model
  2) Measure the accuracy over the test-set for each flip
  3) Mark **Achilles bits** – when the bit flips, it causes the **acc. drop > 10%**

How Vulnerable Are DNNs to a Single Bit-flip?

- **Methodology**
  1) Flip each bit in all parameters of a DNN model
  2) Measure the accuracy over the test-set for each flip
  3) Mark **Achilles bits** – when the bit flips, it causes the **acc. drop > 10%**

- **Quantifying the vulnerability**
  1) **Max. drop** : the maximum acc. drop, observed from a model
  2) **Ratio**. : % of parameters in a model that contains at least one Achilles bit

MNIST Models

Network	Acc.	# Params	Acc. Drop	Ratio
B(ase)	95.71	21,840		
B-Wide	98.46	85,670		
B-PReLU	98.13	21,843		
B-Dropout	96.86	21,840		
B-DP-Norm	97.97	21,962		
L(eNet)5	98.81	61,706		
L5-Dropout	98.72	61,706		
L5-D-Norm	99.05	62,598		

MNIST Models

Network	Acc.	# Params	Acc. Drop	Ratio
B(ase)	95.71	21,840	**98 %**	
B-Wide	98.46	85,670	**99 %**	
B-PReLU	98.13	21,843	**99 %**	
B-Dropout	96.86	21,840	**99 %**	
B-DP-Norm	97.97	21,962	**99 %**	
L(eNet)5	98.81	61,706	**99 %**	
L5-Dropout	98.72	61,706	**99 %**	
L5-D-Norm	99.05	62,598	**98 %**	

- Max. drop $\geq$ **98%** in all models

MNIST Models

Network	Acc.	# Params	Acc. Drop	Ratio
B(ase)	95.71	21,840	**98 %**	**50%**
B-Wide	98.46	85,670	**99 %**	**50%**
B-PReLU	98.13	21,843	**99 %**	**99%**
B-Dropout	96.86	21,840	**99 %**	**49%**
B-DP-Norm	97.97	21,962	**99 %**	**51%**
L(eNet)5	98.81	61,706	**99 %**	**47%**
L5-Dropout	98.72	61,706	**99 %**	**45%**
L5-D-Norm	99.05	62,598	**98 %**	**49%**

- Max. drop $\geq$ **98%** in all models
- **> 45%** of params contain $\geq$ **1 Achilles bit** in all the DNNs

LARGE, COMPLEX DNN MODELS

Dataset	Network	Acc.	# Params	Acc. Drop	Ratio
CIFAR-10	B(ase)	83.74	776K	94 %	46.8%
	B-Slim	82.19	197K	93 %	46.7%
	B-Dropout	81.18	776K	94 %	40.5%
	B-D-Norm	80.17	778K	97 %	45.9%
	AlexNet	83.96	2.5M	96 %	47.3%
	VGG16	91.34	14.7M	99 %	46.2%
ImageNet	AlexNet	79.07	61.1M	100 %	47.3%
	VGG16	90.38	138.4M	99 %	42.1%
	ResNet50	92.86	25.6M	100 %	47.8%
	DenseNet161	93.56	28.9M	100 %	49.0%
	InceptionV3	88.65	27.2M	100 %	40.8%

- Max. drop $\geq$ **98%** in all models
- **> 45%** of params contain $\geq$ **1 Achilles bit** in all the DNNs

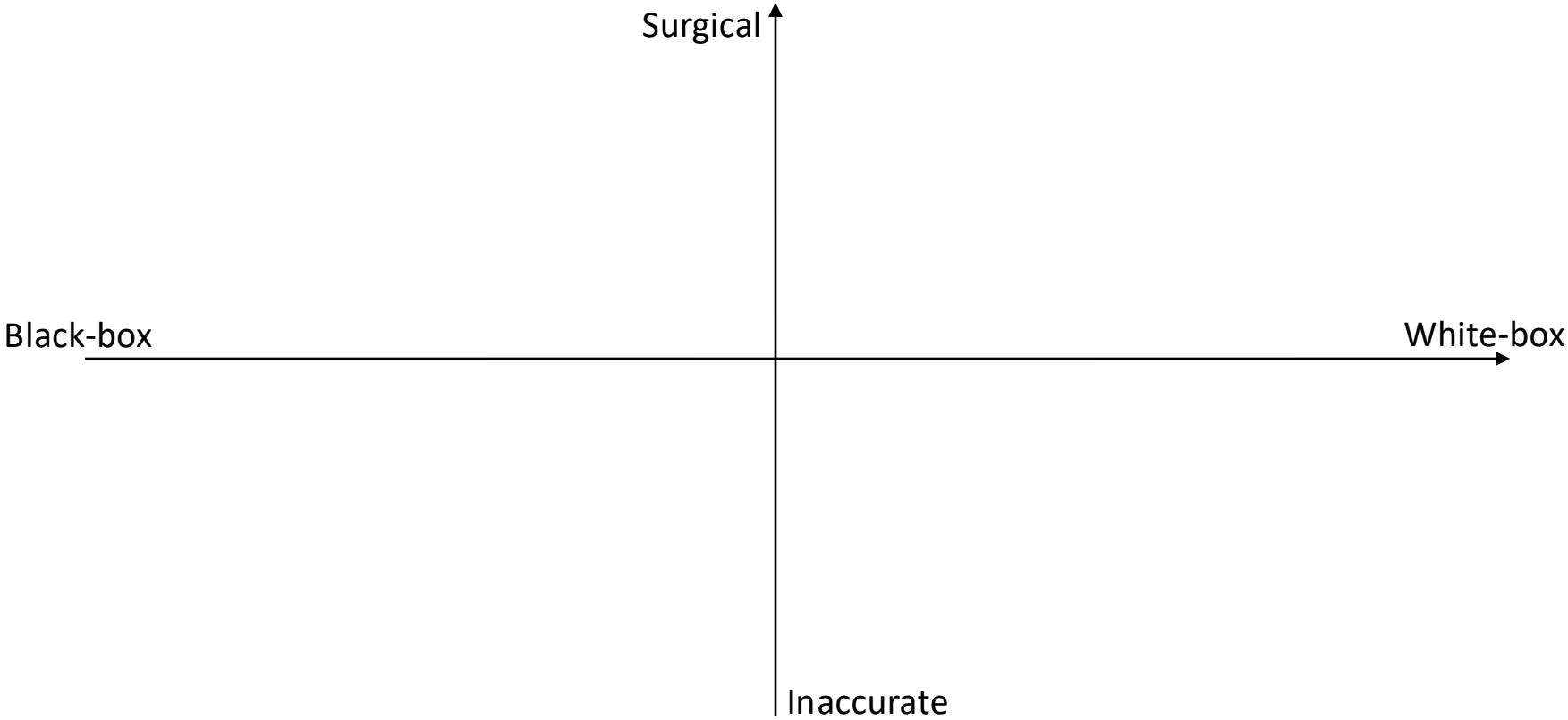
The Vulnerability of DNNs to A Bit-flip Is Prevalent

THREAT MODEL – BIT-FLIP ATTACKER

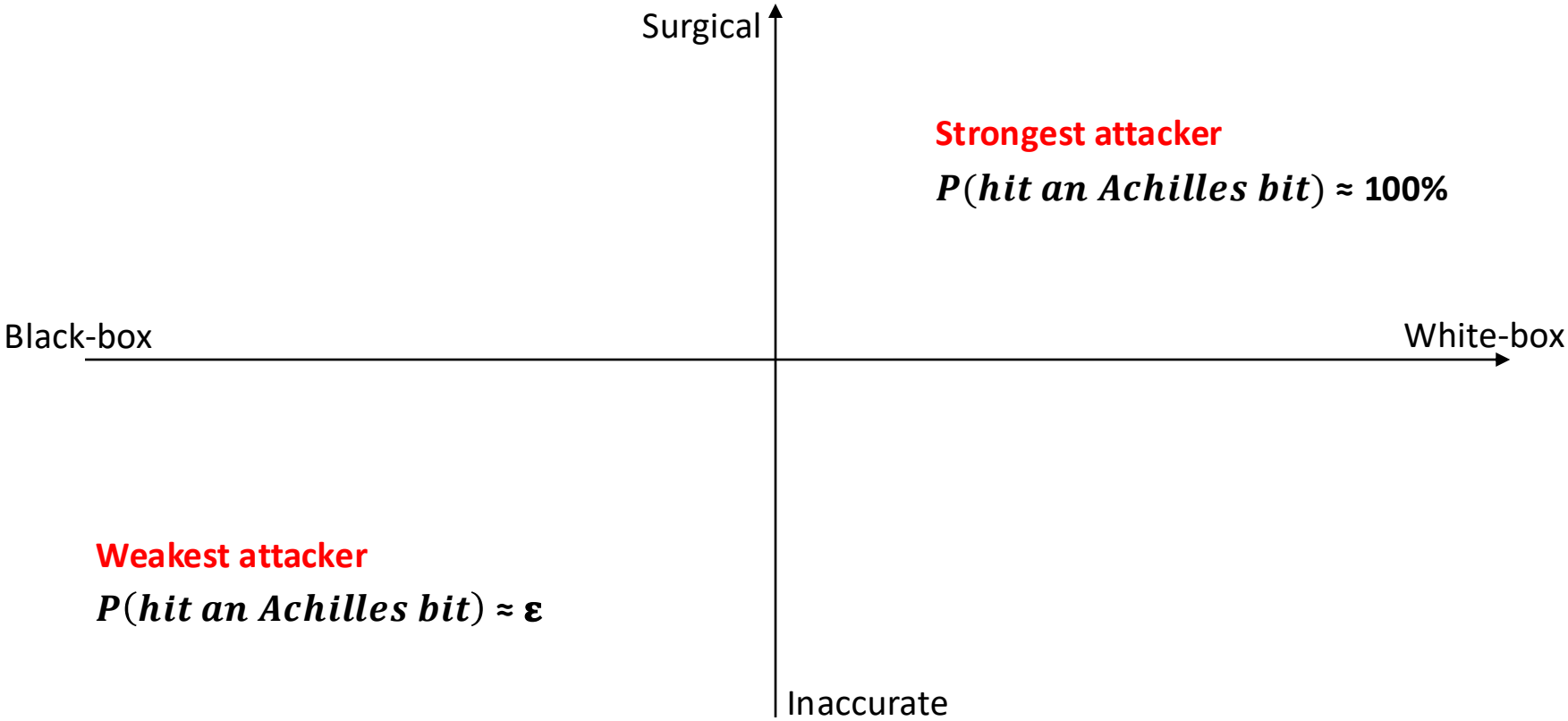
- Capability
  - **Surgical** : can control the location of a bit-flip in memory
  - **Inaccurate**: cannot control the bit-flip location

- Knowledge
  - **White-box**: knows which parameters are vulnerable
  - **Black-box** : has no knowledge of a victim model

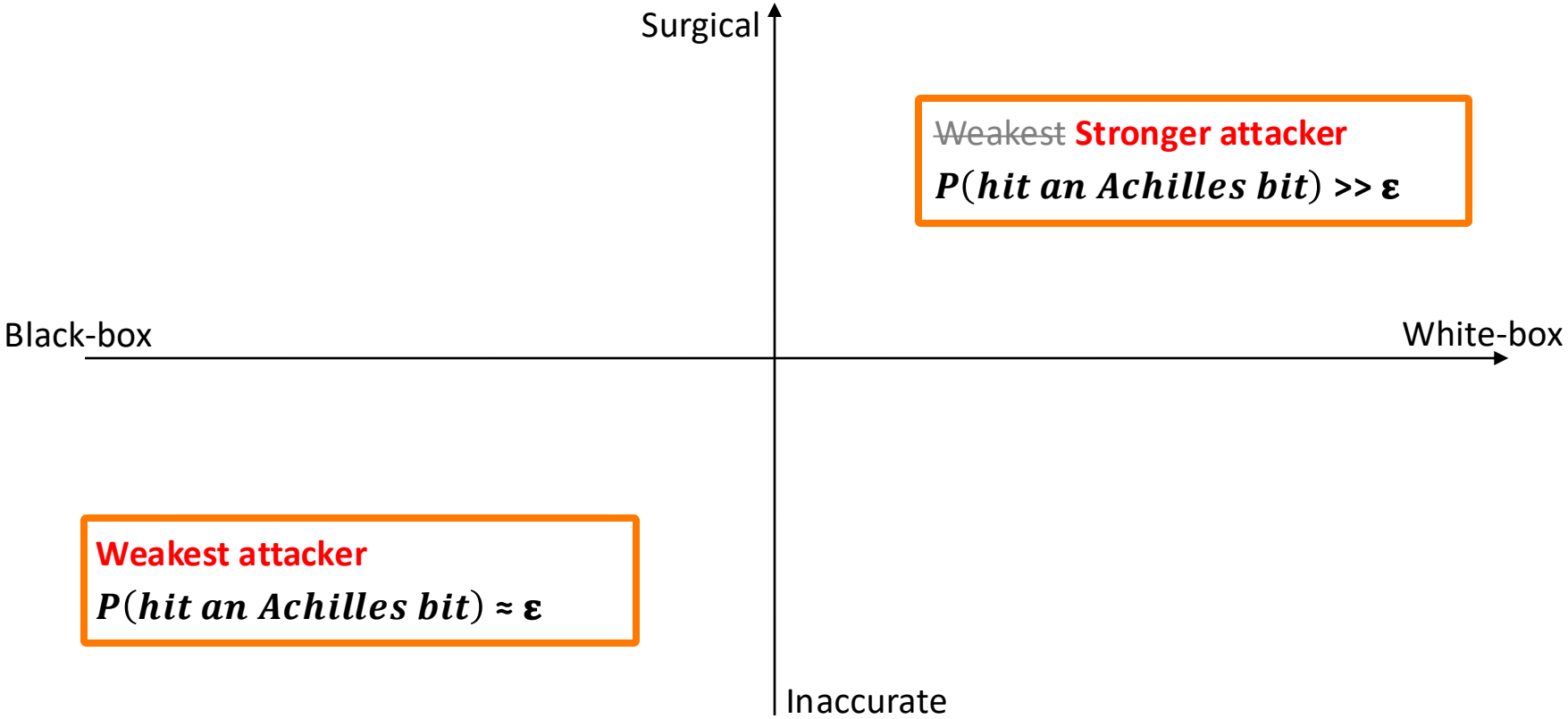
THREAT MODEL – SINGLE-BIT ADVERSARY



THREAT MODEL – SINGLE-BIT ADVERSARY

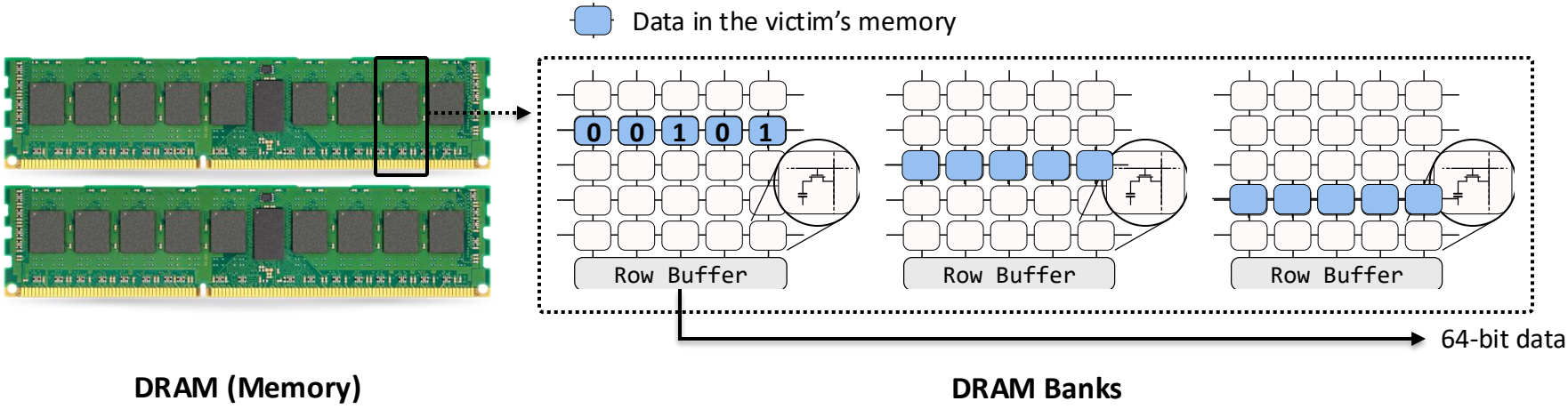


THREAT MODEL – IF THE ADVERSARY CAN FLIP MULTIPLE BITS?



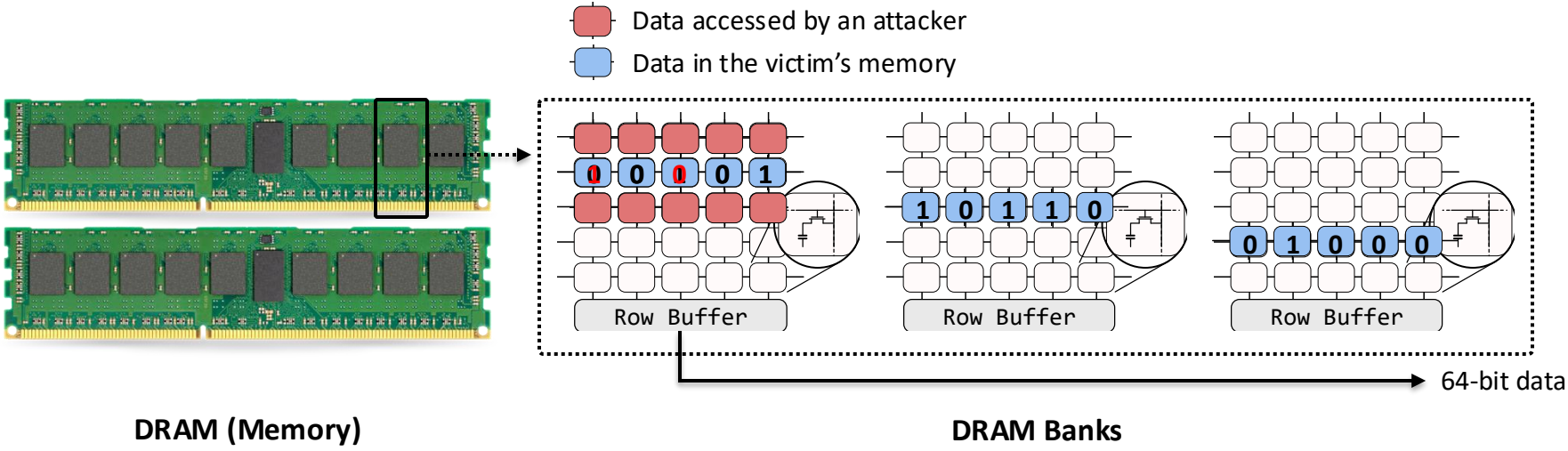
PRACTICAL HARDWARE ATTACK – ROWHAMMER

- Rowhammer attacks
  - Single-bit corruption primitives in DRAM-level
  - Software-induced hardware fault attack



PRACTICAL HARDWARE ATTACK – ROWHAMMER

- Rowhammer attacks
  - Single-bit corruption primitives in DRAM-level
  - Software-induced hardware fault attack
  - Cross-VM: attacker only requires a co-located VM



EVALUATION

- **MLaaS scenario**
  - **Victim** : runs an off-the-shelf model (VGG16) in a VM
  - **Attacker** : runs Rowhammer attacks against the victim's VM

- **Rowhammer** (Hammertime[1] DB)
  - Explore Rowhammer attacks systematically on **12 different DRAM chips**
  - Experiments:
    - ***300*** *experiments*: ***25*** *runs* × *each of 12 DRAM chips*
    - ***7500*** *bit-flips* : ***300*** *cumulative bit-flips* × ***300*** *experiments*

[1]Tartar et al., *Defeating Software Mitigations against Rowhammer: A Surgical Precision Hammer*, RAID'18

EVALUATION

- **Results**
  - The weakest attacker can inflict **severe damage** to the victim system
    - On average, **62%** of the attacks cause the **acc. drop > 10%**
    - The time it takes to cause the acc. drop is **< few minutes**
  - Our attack is **inconspicuous**
    - Only 6 program crashes (**0.08%**) were observed over 7500 bit-flip attempts

TAKEAWAYS

- DNNs are ***not*** resilient to worst-case param. perturbations
  - All DNNs have a bit whose flip causes the accuracy drop up to 100%
  - 40-50% of all parameters in a model are vulnerable

- **The vulnerability of DNNs to fault attacks is under-studied**
  - One can inflict the vulnerability with ***weaker attacks***, *e.g.*, blind Rowhammer
  - The attacker can launch this attack in a practical setting, *e.g.*, in the cloud

- **We need solutions from both systems and ML**
  - **Systems:** defenses that prevent flipping a specific-bit are not sufficient
  - **ML:** future work is required to build DNNs robust against new attacks

Thank You!

Sanghyun Hong

https://secure-ai.systems/courses/Sec-Grad/current

SAIL
Secure AI Systems Lab