360 Overview of Semiconductor from AI and Security Perspective

Claudionor N. Coelho Jr, PhD/MBA
Chief AI Officer / SVP of Engineering
May 2023
## Agenda

1. New Challenges in Semiconductor Manufacturing
2. Data is the New Oil
4. True Zero Trust™ with Unified Server
5. With Information Highway, We Can Embrace AI
New Challenges in Semiconductor Manufacturing

- Chip-level scale integration
- New packaging: 2.5D and 3D Chiplets
- Chip-scale sensors requiring
- System quality challenge
New Challenges in Semiconductor Manufacturing

Cores That Don’t Count*

Peter H. Hochschild  Paul Turner  Jeffrey C. Mogul  Rama Krishna Govindaraju  Parthasarathy Ranganathan  David E. Culler  Amin Vahdat

Proc. 18th Workshop on Hot Topics in Operating Systems (HotOS 2021)

Abstract

We are accustomed to thinking of computers as fail-stop, especially the cores that execute instructions, and most system software implicitly relies on that assumption. During most of the VLSI era, processors that passed manufacturing tests and were operated within specifications have insulated us from this fiction. As fabrication pushes towards smaller feature sizes and more elaborate computational structures, and as increasingly specialized instruction-silicon pairings are introduced to improve performance, we have observed ephemeral computational errors that were not detected during manufacturing tests. These defects cannot always be mitigated by techniques such as microcode updates, and may be correlated to specific components within the processor, allowing small code changes to effect large shifts in reliability. Worse, these failures are often "silent": the only symptom is an erroneous computation.

We refer to a core that develops such behavior as “mercurial.” Mercurial cores are extremely rare, but in a large fleet of servers we can observe the correlated disruption they cause, often enough to see them as a distinct problem – one that will require collaboration between hardware designers, processor vendors, and systems software architects.

*https://research.google/pubs/pub50337/
We Will Need Data from All Sources to Help with New Challenges

Open Solutions Ecosystem

ACS Products & Services
- ACS TE-Cloud
- ACS DPT

Customer Applications
- Third Party Database

Third Party Applications
- APC
- DEF
- DPAT
- Chiplets

ACS Real-Time Data Infrastructure
ACS Nexus (Data Access) + ACS Edge (Edge Computing) + ACS Unified Server (True Zero Trust™ Security)

New Solutions Drive Infrastructure Requirements
New Infrastructure Enables Solution Innovation

Customer Applications
- Data
- Data
- Data
- Data
- Data

Third Party Applications
- Data
- Data
- Data
- Data
- Data

ACS Products & Services
- Data
- Data
- Data
- Data
- Data

ACS Real-Time Data Infrastructure

PSV
WAT
WS
FT
BURN-IN
SLT
Should I Worry About Security?
Protection

- Security updates
- Removing unnecessary drivers
- Closing unnecessary ports
- Limiting access to systems
- Using firewalls, anti-virus, malware, spyware protection applications
- Encryption
- Password management
- Intrusion prevention systems (IPS)
- Intrusion detection systems (IDS)

Result of protection of my home against known threats

We’ve blocked 522 threats to your network
Since January, 2022

- Spam (460) 89%
- Phishing (9) 2%
- Attacks (27) 6%
- Fraud (5) 1%
- Malware (21) 5%
Rubrik CEO: Cyber Attacks Are Inevitable

Rubrik CEO Bipul Sinha joins Ed Ludlow to discuss the state of cybersecurity even as the Pentagon leak investigation is still ongoing, what companies can do to better protect themselves and why the company is waiting for the market to "be ready" before going public. (Source: Bloomberg)
Digression: Let’s Talk About the ISS
The ISS has shields called Whipple bumpers. They are multi-layered with spaces between the layers. The intent is that impact with a layer will both slow and hopefully break apart the projectile, so that by the time it gets to the bottom layer it is no longer harmful.
Eventually, debris will go through, and air leaks will need to be contained.

https://www.esa.int/ESA_Multimedia/Images/2015/03/Space2_on_Columbus
… this shows is just **not enough to protect** …

… we also **need to contain** …
Usually, We Think about Test Cell
… But Test Cell is Included in a Test Floor …
... And Test Floor is Surrounded by Other Equipment
Target’s 2013 Data Hack – What Happened and How?

Four years ago, the Target Corporation found itself in the crosshairs of dedicated hackers that resulted in one of the largest attacks on a retail chain in recent years. As a result, 40 million Target customers’ records were stolen, and upwards of 70 million further records were compromised. The fallout of this incident was a nationwide lawsuit against the corporation, involving 47 states as well as the District of Columbia. As of May 2017, this legal battle was settled with Target paying out $18.5 million in settlement claims. Apart from the legal costs, Target disclosed that the hack cost the company an estimated $202 million.

Target Corporation claims to yield impressive data security programs that are similarly used by Federal agencies, such as the CIA and the Pentagon, and yet a breach of this magnitude was still successful. Compared to the immediate aftermath of the incident, today we know considerably more about the nature of the hack and, more importantly, how it could have been prevented.

The method of the hack itself was relatively unorthodox, yet clearly effective. Instead of attempting to breach Target’s security measures protecting its databases, malware was instead allowed to infiltrate the company via its point of sale machines. The investigation into this attack discovered that the origin of the malware came from a phishing email that was sent to employees of Fazio Mechanical, an HVAC firm that was hired by Target, and was subsequently downloaded (likely by accident) by an employee.

... And Test Floor is Surrounded by Other Equipment

- Printer
- Firewall
- Router
- Software
- Computer

IoT Coffee Pot
Nearly A Million Printers At Risk Of Attack, Thousands Hacked To Prove It

Lee Mathews  Senior Contributor  
Observing, pondering, and writing about tech. Generally in that order.

Nearly A Million Printers At Risk Of Attack, Thousands Hacked To Prove It (forbes.com)
... And Test Floor is Surrounded by Other Equipment
Attacks Can Target Software Update Servers …

Possible channel

Alan Woodward, a computer scientist from the University of Surrey, said: The ironic thing about this situation (if it proves to be the case) is that we always advise users to keep their software up to date, ideally using automated updates.

“However, it assumes hackers can’t take over the update process and misuse it.

“This process is normally a very tightly controlled process, so this is unusual.

“I can imagine many vendors are now triple-checking to make sure they don’t end up being an attack vector.”

He said that it showed *hackers will probe every possible channel* to find a route into systems.

“As users there isn’t a lot we can do as we are in the hands of the software vendors.”

[Tax software blamed for cyber-attack spread - BBC News](https://www.bbc.com)
... And Test Floor is Surrounded by Other Equipment

IoT Coffee Pot

Printer

Firewall

Router

Software

Computer
Microsoft warns of North Korean crew posing as LinkedIn recruiters

State-sponsored ZINC allegedly passes on malware-laden open source apps

Laura Dobberstein  Fri 30 Sep 2022  05:53 UTC

Microsoft has claimed a North Korean crew poses as LinkedIn recruiters to distribute poisoned versions of open source software packages.

The state-sponsored group has been around since 2009 and was allegedly behind the 2014 attack on Sony Pictures in retaliation for the controversial Seth Rogen comedy The Interview.

Dubbed "ZINC", the threat actors have previously run long-term phishing schemes targeting media, defence and aerospace, and IT services organizations in the US, UK, India, and Russia.

Starting in June of this year, ZINC relied on social engineering tactics: contacting targets on LinkedIn and claiming to be a recruiter, establishing trust with targets, and switching communications to WhatsApp where they delivered shellcode from the ZetaNile malware family.

https://www.theregister.com/2022/09/30/microsoft_north_korea_zinc_threat/
... And Test Floor is Surrounded by Other Equipment
Dozens of Netgear routers can easily be hacked — what to do right now [updated]

By Paul Wagenseil last updated March 30, 2021

Nearly 80 Netgear router models could be hacked over the Internet, but workarounds are trickling out

Hello again, old friend. We've met here before. (Image credit: Netgear)

How to Bypass a Firewall or Internet Filter

Written by Jack Lloyd
Last Updated: November 25, 2022  Tested

This wikiHow teaches you how to view blocked websites or content on a restricted computer, as well as on a mobile item if you're using a Virtual Private Network (VPN).

https://www.wikihow.com/Bypass-a-Firewall-or-Internet-Filter
... Some Test Floors Require Cloud Connection ...
What Your Firewall Won’t Protect You From

**INSIDER THREATS**

Unfortunately, there are plenty of enemies within the gate, sometimes acting intentionally, other times inadvertently creating vulnerabilities.

For example, disgruntled employees can pose a significant threat to your electric cooperative. Already behind the firewall and familiar with the system, they could mass delete files or intentionally install a virus.

A less malevolent (but no less dangerous) threat is shadow IT. This is any software, applications, or firewall exceptions team members install or enable, usually to eliminate steps in a process or make their jobs easier. Though they may not have malicious intent, it’s not uncommon for employees to weaken a firewall by adding an excess number of exceptions to it so they can complete a task more quickly. Some of the applications an employee may innocently install may create backdoor access to the co-op’s system.

Not every piece of malware gets into your co-op’s system through a firewall, either. Computer viruses and other attacks can just as easily get into your system through a USB stick someone has brought in from outside. Again, this may even happen unintentionally.

**DATA LEAKAGE**

Firewalls may also not be able to protect your co-op from data leakage, or the unauthorized transmission of data leaving your electric co-op. Normally, data leakage happens through the web or email but it is possible to cause data to leak through hardware.

State-sponsored organizations have been known to purposely sabotage equipment so that it leaks data from the organization. Often, this happens in such small amounts even a firewall won’t detect something is wrong because the device is supposed to send data as part of its normal operations. Imagine buying a brand-new switch, manufactured overseas, only to discover much later it had been sending sensitive data about you the entire time the leak was unknown. In this situation, even an advanced firewall may not be helpful.

Most data leakage, however, is done using malware specifically designed to slowly leak information out of your cooperative.

Regardless of how it is occurring, data leakage is a very real threat that must be addressed. Simply relying on a firewall to keep your data safe will not keep your electric cooperative safe from this specific kind of cyberattack.

https://skyhelm.com/cyberattacks-your-firewall-wont-protect-you-from/
Staging a Data Leakage

secret message

hidden_msg = "***
Integrated Annual Report 2022
Advantest has published our Integrated Annual Report 2022. In addition to financial information, the report includes non-financial information on our Grand Design, vision, strategy, results, governance, and sustainability. It provides a comprehensive snapshot of our corporate value and prospects for medium- to long-term growth."***

measurements

[ 2.2393817 -0.8189581 1.707974  0.91654325  0.53305906  0.22747241 
-0.5253728 -1.4496815  0.62295085 -0.4778164 ]

-0.4778164

Malicious agent can change noise level of measurements and transmit confidential information

Insider threat
Data leakage
Web-proxy
Cyber Attacks are inevitable. Protection is not enough any longer. Containment is the only way to secure you have limited leakage.
True Zero Trust™ Environment for Test Floor is a Must…

• Provides containment environment for Test Floor
• Assumes facility has been compromised upward or downward
• Shared ownership of security by OSAT and Fabless
• Enables packet inspection and logging of transactions
• Provides transaction-level security
• Encrypted traffic to Unified Server
• Third party to provide full SBOM of packages to detect known threats
ACS Real-Time Data Infrastructure (RTDI)

Customer

Data Scientist
Develops and tests AI / ML workloads

Test Engineer
Develops and releases test program

Key Features:

• Communication backplane for Test Floor
• Edge level computing facility
• Floor level True Zero Trust™ with Unified Server Container hub

Test Facility

Test Cell

ACS Container Hub

Management of AI/ML container images

Unified Server Container Mirror

On-prem copy from cloud

Host Controller

ACS Nexus Software

ACS Edge Server
x86 server with optional GPU

10Gb/s Encrypted Ethernet
ACS Real-Time Data Infrastructure Securely Supports...

• Data feed forward / backward with full multi-site integration
• Outlier detection
• Adaptive test
• Predictive modeling
• Optimization of heavy computation test
What Can You do with Edge and Unified Server Infrastructure…

The Brick

Mobile phones used to be bought for phone calls…
What Can You do with Edge and Unified Server Infrastructure…

Mobile phones used to be bought for phone calls…

…nowadays, mobile phones can also be used for phone calls…

Global mobile data traffic Exabytes per month

Source: Ericsson Mobility Report
Embracing AI New Technologies
Which New AI Technologies Can Be Used With More Data

New chip technologies require new measurements, relying in multi-dimensional data

Large Language Models are creating vast new opportunities in all domains
Thermal Profile of Dies

https://semiengineering.com/a-high-level-how-to-guide-for-effective-chip-package-thermal-co-design/
What’s an LLM (Large Language Model)?

- Very large number of parameters
- Trained on a very large corpus (petabytes)
- Trained for several months and with several restarts and with a hefty electricity bill
- People do not know how information is stored because of its size

Figure 1: Exponential growth of number of parameters in DL models

ChatGPT

GPT-3

GPT-4

175,000,000,000

100,000,000,000,000
Using LLMs (Large Language Models)

LLMs can comprehend and link topics and it has “SOME” understanding of math, but ChatGPT does not understand new developments as it is not connected to the Internet.

Large Language Models are Zero-Shot Reasoners

Takeshi Kojima
The University of Tokyo
t.kojima@e2lab.i.u.tokyo.ac.jp

Shixiang Shane Gu
Google Research, Brain Team

Machal Reid
Google Research

Yutaka Matsuo
The University of Tokyo

Yusuke Iwasawa
The University of Tokyo

Large Language Models Are Reasoning Teachers

Namgyu Ho, Laura Schmid, and Se-Young Yun
KAIST
{itsnamgyu, laura.schmid, yunseyoung}@kaist.ac.kr

solve x**2 -5 * x + 3 = 0 for x

This is a second-degree equation and can be solved using the quadratic formula:

\[ x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \]

Where a = 1, b = -5, c = 3 and the equation is written as:

\[ x^2 - 5x + 3 = 0 \]

So, plugging in the values, we get:

\[ x = \frac{(-5) \pm \sqrt{((-5)^2 - 4 \cdot 1 \cdot 3)}}{2 \cdot 1} \]

\[ x = \frac{(5 \pm \sqrt{25 - 12})}{2} \]

\[ x = \frac{(5 \pm \sqrt{13})}{2} \]

Therefore, the solutions are:

\[ x = (5 + \sqrt{13}) / 2 \text{ and } x = (5 - \sqrt{13}) / 2 \]
Using LLMs (Large Language Models)

Search Engines list documents. LLMs summarize knowledge.

Here is some sample Python code to test the function `func(a,b):

```python
def test_func(a, b):
    expected = a + b
    actual = func(a, b)
    assert expected == actual

test_func(1, 2)
test_func(3, 4)
```

This code will call the function `func(a,b)` with two different sets of inputs and check that the output matches the expected result. If the output does not match, an `AssertionError` will be raised.
Using LLMs (Large Language Models)

LLMs don’t know when it makes mistakes, so it should be considered as an exoskeleton
Conclusions

Over the next few years, we can anticipate a significant shift in the types of applications being developed, moving away from traditional statistical machine learning and towards more sophisticated autonomous or semi-autonomous agents that can automate testing activities.

In order to effectively safeguard the valuable assets and intellectual property of OSAT and Fabless organizations, containment is necessary. ACS Real-time Data Infrastructure offers a highly secure containment system called True Zero Trust™.

Through its innovative design, ACS real-time infrastructure establishes a cutting-edge paradigm that allows for the creation of secure data highways and paves the way for building novel applications with enhanced security.
Questions?