



Principled Secure Processor Design

CHRISTOPHER W. FLETCHER

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN





Thank you

To my students and collaborators ©

Jiyong Yu, Mohamad El Hajj, Lucas Hsiung, Mengjia Yan, Namrata Mantri, Artem Khyzha, Adam Morrison, Josep Torrellas, Po-An Tsai, Andres Sanchez, Daniel Sanchez



How good is process isolation nowadays, anyway?



Computing Abstractions Today (ideal)

r = P(s)

secrets are red

P, Encrypt(s)

Processor (black box)

Encrypt(r)

OS: protects computation P(s)

Crypto: protects data in transit (s, r)

→ No unauthorized party learns s



Computing Abstractions Today (really)

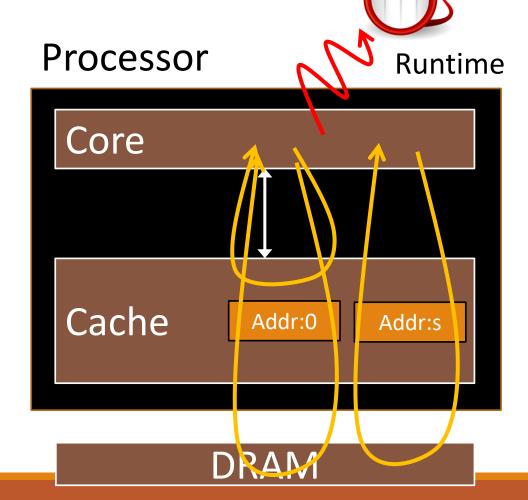
```
// s = bool
void P(secret s) {
  load(0);
  load(s*BLOCK_SZ);
}
```

Case 1: s = 0

Cache: miss, hit

 \rightarrow P is fast

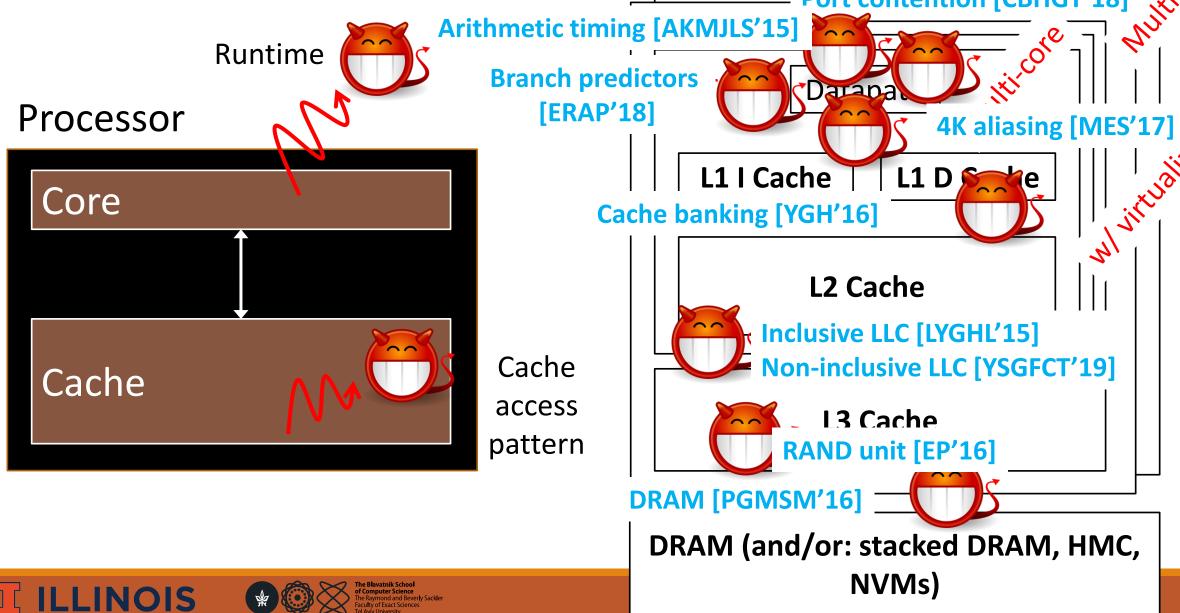
Case 2: s = 1 miss, miss → P is slow







Microarchictural Side/Covert Channels Everywhere Port contention [CBHGT'18]



Worse: attacks can enable "read gadgets"

```
Unsafe:
```

```
void P(secret s) {
  load(0);
  load(s*BLOCK_SZ); }
Safe:
void P(secret s) {
  load(0);
  load(0*BLOCK SZ);
```

load(1*BLOCK SZ); }

Read gadget:

bool ← read(addr a)

- Attacker controls a
- Leaks P's memory bit by bit

This talk:

Principled, low-overhead defenses against microarchitectural attacks**

** FOCUSING TODAY ON SPECULATIVE EXECUTION ATTACKS



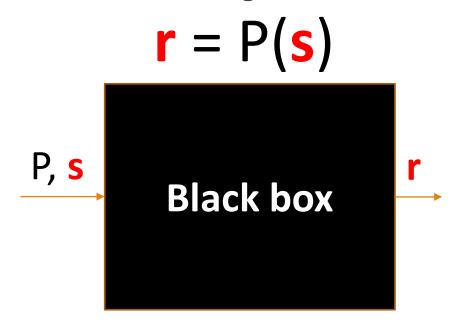


Principled, Low-overhead

Many uarch side/covert channel (cache, predictors, etc.)



Want: some clean security definition E.g.,



i.e., secure given any uarch side/covert channels



Principled, Low-overhead

Obviously.

But not at the expense of clean security.





A lattice model of secure information flow; Dorothy E. Denning, CACM 1976



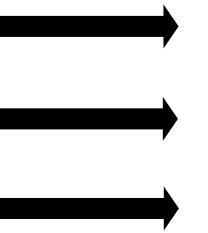


Security + Low-overhead



Classified Unclassified

Unclassified —— Classified Unclassified — Unclassified



Classified —— Classified 🕒



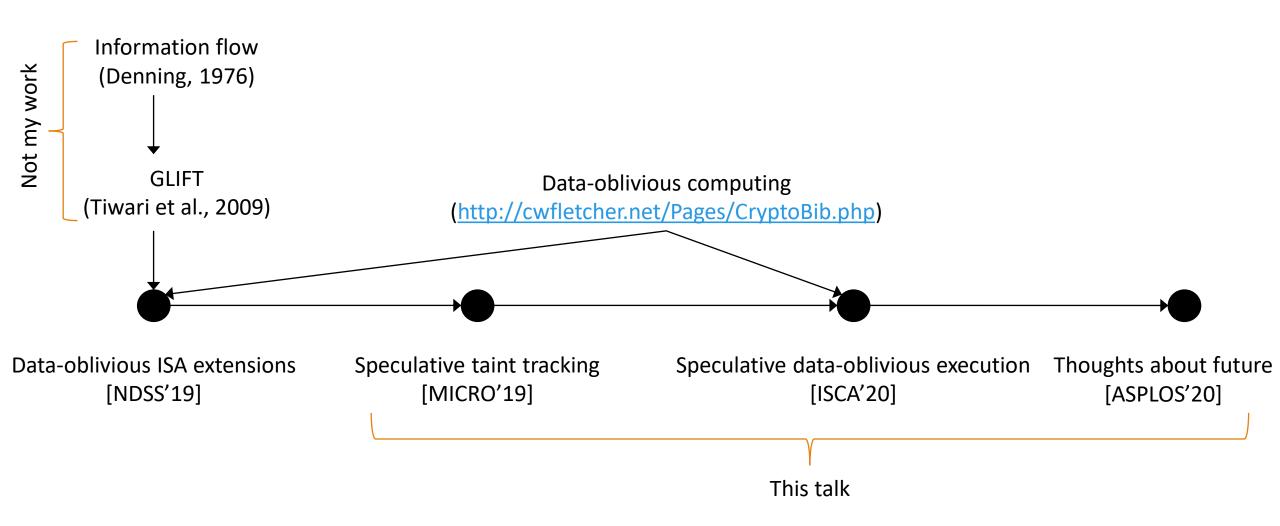








This talk



Part 1: Speculative Taint Tracking (STT)

COMPREHENSIVE PROTECTION FOR SPECULATIVE DATA





Speculative Execution Attacks*

```
// Spectre Variant 1

if (addr < N) { // speculation

    // access instruction
    spec_val = load [addr];

    // covert channel
    load [spec_val];</pre>
```

```
Speculation starts
```

*: Kocher et al.; "Spectre Attacks: Exploiting Speculative Execution", SP'19.









```
Speculation starts

// Spectre Variant 1

if (addr < N) { // speculation

// access instruction
spec_val = load [addr];

// covert channel
load [spec_val];
}</pre>
Speculative access instruction**

accesses secret
```

*: Kocher et al.; "Spectre Attacks: Exploiting Speculative Execution", SP'19.

**: Kiriansky, Vladimir, et al.; "DAWG: A defense against cache timing attacks in speculative execution processors." MICRO'18.









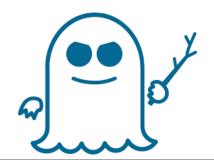
^{**:} Kiriansky, Vladimir, et al.; "DAWG: A defense against cache timing attacks in speculative execution processors." MICRO'18.





^{*:} Kocher et al.; "Spectre Attacks: Exploiting Speculative Execution", SP'19.



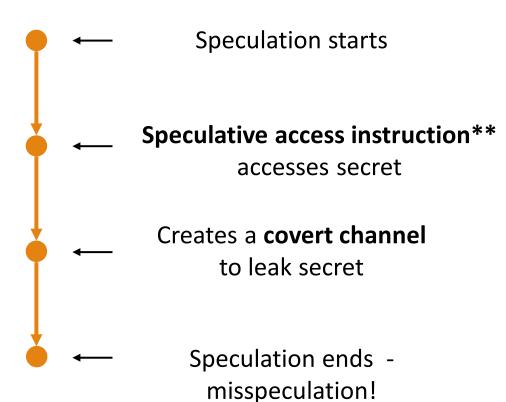


```
// Spectre Variant 1

if (addr < N) { // speculation

addr = // access instruction
    spec_val = load [addr];

// covert channel
    load [spec_val];
}</pre>
```

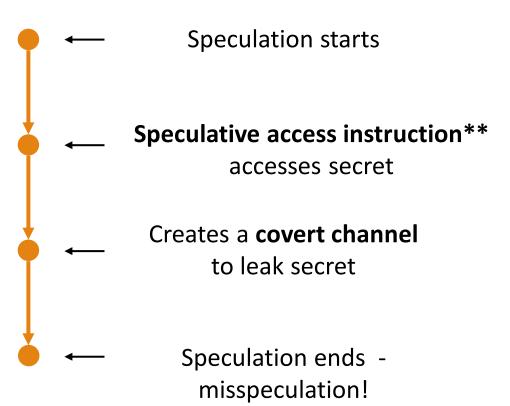


*: Kocher et al.; "Spectre Attacks: Exploiting Speculative Execution", SP'19.

^{**:} Kiriansky, Vladimir, et al.; "DAWG: A defense against cache timing attacks in speculative execution processors." MICRO'18.







*: Kocher et al.; "Spectre Attacks: Exploiting Speculative Execution", SP'19.



^{**:} Kiriansky, Vladimir, et al.; "DAWG: A defense against cache timing attacks in speculative execution processors." MICRO'18.



Speculative Execution Attacks







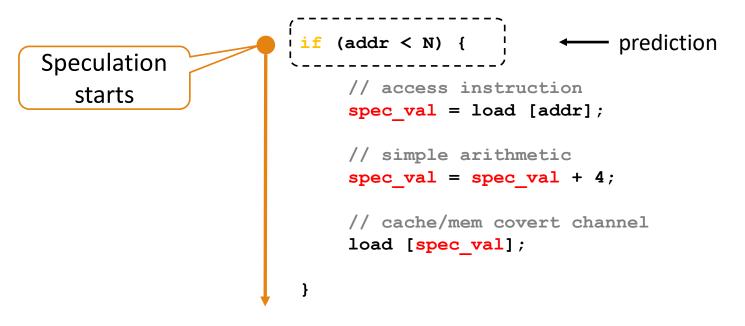
```
if (addr < N) {
    // access instruction
    spec_val = load [addr];

    // simple arithmetic
    spec_val = spec_val + 4;

    // cache/mem covert channel
    load [spec_val];</pre>
```

Creates a covert channel?	Input operand is a secret?	Requires protection?





Creates a covert channel?	Input operand is a secret?	Requires protection?

"Sufficient for security: prevent secrets from reaching covert channels"

Speculation starts

```
if (addr < N) {
    // access instruction
    spec_val = load [addr];

    // simple arithmetic
    spec_val = spec_val + 4;

    // cache/mem covert channel
    load [spec_val];</pre>
```

Creates a covert channel?	Input operand is a secret?	Requires protection?
Yes	No	No

```
Speculation starts
```

```
if (addr < N) {
    // access instruction
    spec_val = load [addr];

    // simple arithmetic
    spec_val = spec_val + 4;

    // cache/mem covert channel
    load [spec_val];</pre>
```

Creates a covert channel?	Input operand is a secret?	Requires protection?
Yes	No	No
No	Yes	No

"Sufficient for security: prevent secrets from reaching covert channels"

Speculation starts

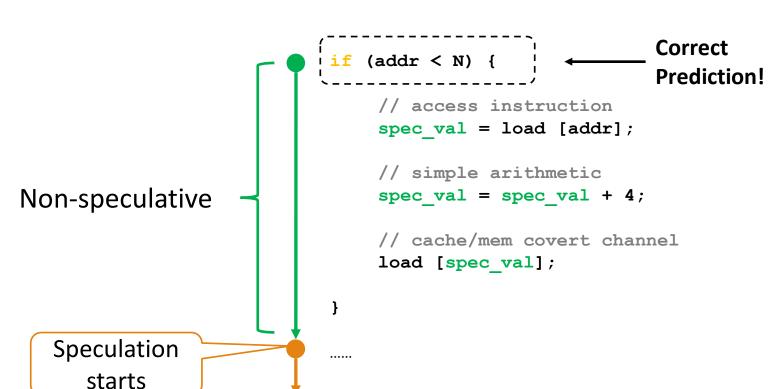
```
if (addr < N) {
    // access instruction
    spec_val = load [addr];

    // simple arithmetic
    spec_val = spec_val + 4;

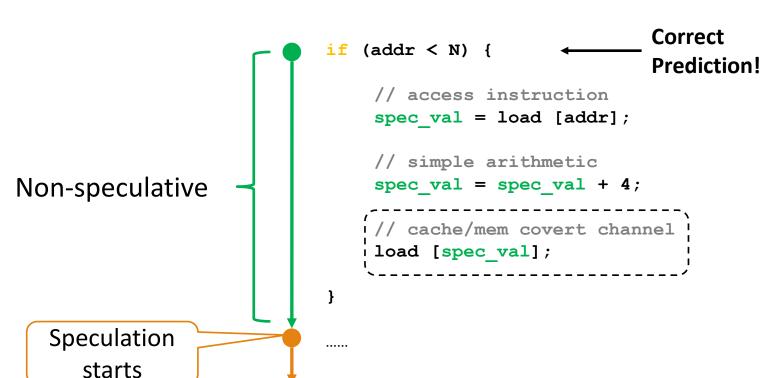
    // cache/mem covert channel
    load [spec_val];
}</pre>
```

Creates a covert channel?	Input operand is a secret?	Requires protection?
Yes	No	No
No	Yes	No
Yes	Yes	Yes





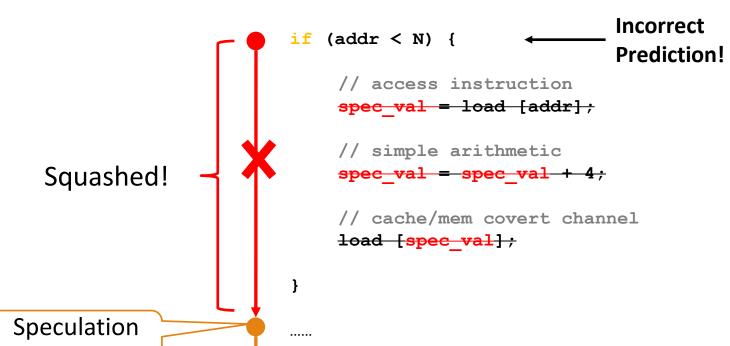
Creates a covert channel?	Input operand is a secret?	Requires protection?
Yes	No	No
No	No	No
Yes	No	No

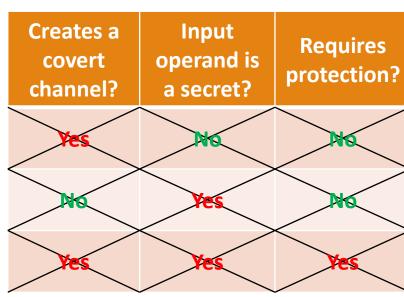


Creates a covert channel?	Input operand is a secret?	Requires protection?
Yes	No	No
No	No	No
Yes	No	No



"Sufficient for security: prevent secrets from reaching covert channels"





starts

Secret (speculatively accessed data)



Covert channels



Secret (speculatively accessed data)



Covert channels

Security definition:

Arbitrary speculative execution can only leak retired register file state.





Secret (speculatively accessed data)



Covert channels







Secret (speculatively accessed data)



Covert channels





What are the covert channels?



Secret (speculatively accessed data)



Covert channels





What are the covert channels?

A new classification to understand covert channels in speculative machines



Secret (speculatively accessed data)



Covert channels





What are the covert channels? ——

A new classification to understand covert channels in speculative machines



How to identify all the secrets?

Speculative Taint Tracking

Secret (speculatively accessed data)



Covert channels





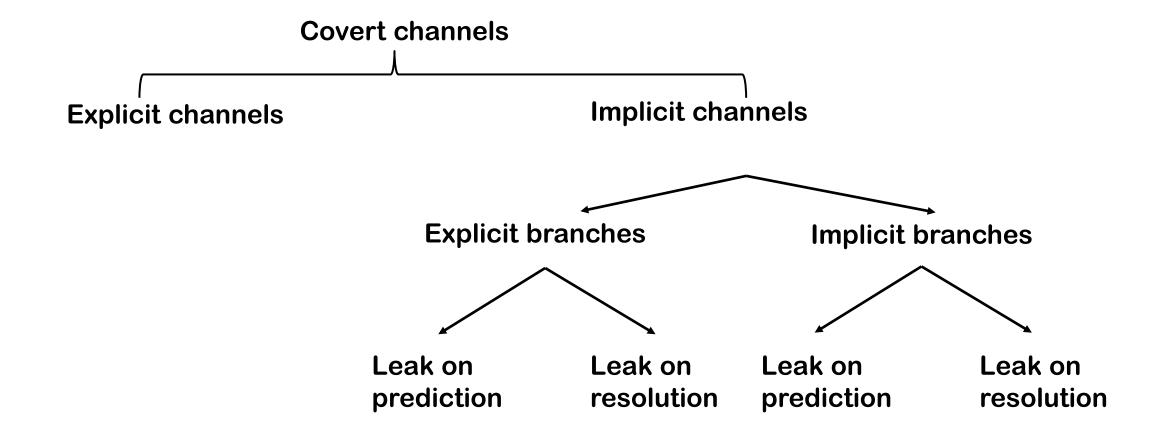


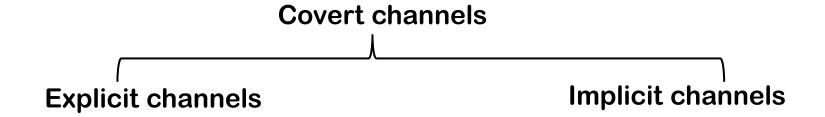
How to identify all the secrets?

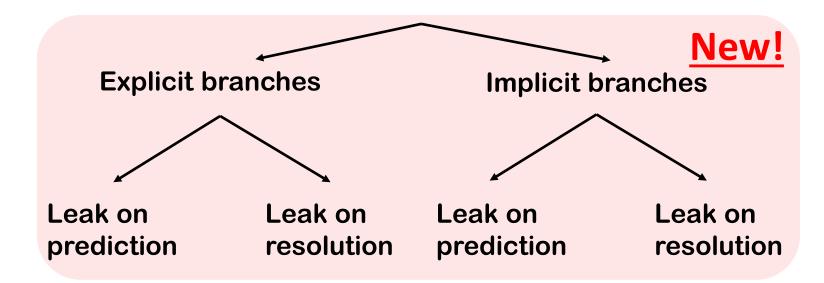
A new taint/untaint mechanism to track secrets in hardware

A Classification of Covert Channels in HW











Covert channels

Explicit channels:

Secret inputs are directly leaked by operand-dependent hardware resource usage

load [secret];

Covert channels

Explicit channels:

Secret inputs are directly leaked by operand-dependent hardware resource usage

Examples:

memory loads data-dependent arithmetic





Covert channels

Explicit channels:

Secret inputs are directly leaked by operand-dependent hardware resource usage

Examples:

memory loads data-dependent arithmetic

Implicit channels:

Secret inputs are indirectly leaked by how (or that) one or several instructions execute

```
secret = load [addr];
if (secret == 1)
    load [0x00];
```



Covert channels

Explicit channels:

Secret inputs are directly leaked by operand-dependent hardware resource usage

Examples:

memory loads data-dependent arithmetic

Implicit channels:

Secret inputs are indirectly leaked by how (or that) one or several instructions execute

```
secret = load [addr];
if (secret == 1)
    load [0x00];
```



Covert channels

Explicit channels:

Secret inputs are directly leaked by operand-dependent hardware resource usage

Examples:

memory loads data-dependent arithmetic

Implicit channels:

Secret inputs are indirectly leaked by how (or that) one or several instructions execute

Examples:

branch/jump instructions



Covert channels

Explicit channels:

Secret inputs are directly leaked by operand-dependent hardware resource usage

Examples:

memory loads data-dependent arithmetic

Implicit channels:

Secret inputs are indirectly leaked by how (or that) one or several instructions execute

Explicit branches

Examples:

Branch/jump instructions







Covert channels

Explicit channels:

Secret inputs are directly leaked by operand-dependent hardware resource usage

Examples:

memory loads data-dependent arithmetic

Implicit channels:

Secret inputs are indirectly leaked by how (or that) one or several instructions execute

New

Explicit branches

Examples:

Branch/jump instructions

Leak on prediction

Leak on resolution





Secrets are red

Non-secrets are green

Attacker can see sequence of memory accesses (to L1 cache)

```
secret = load [0x00];
if (secret == 1)
  load [0x01];
else
  load [0xFF];
```

```
Case 1 (secret == 1): Attacker sees [0x00, 0x01]
Case 2 (secret == 0): Attacker sees [0x00, 0xFF]
```

```
... ...
if ( secret )
... ...
if ( public )
    load [0x00];
else
    load [0x10];
```



Cause:

The predictor state becomes a function of secret

```
... ...
if ( secret )
... ...
if ( public )
    load [0x00];
else
    load [0x10];
```



Cause:

The predictor state becomes a function of secret

```
Resolve and update
if (secret)
branch predictor

if (public)
load [0x00];
else
load [0x10];
Branch Predictor Unit (BPU)
```



Cause:

The predictor state becomes a function of secret

```
if ( secret )
if ( public )
load [0x00];
else
load [0x10];

Branch Predictor Unit (BPU)
```



Covert channels

Explicit channels:

Secret inputs are directly leaked by operand-dependent hardware resource usage

Examples:

memory loads data-dependent arithmetic

Implicit channels:

Secret inputs are indirectly leaked by how (or that) one or several instructions execute

Explicit branches Examples:

Branch/jump instructions

Leak on prediction resolution







Explicit Branches @ Resolution

```
if (secret) {
    y++;
}
z = load [0x00]
```



Explicit Branches @ Resolution

Cause:

The resolution of a mis-speculation triggers a pipeline squash and alternation of control flow

```
if (secret) {
    y++;
}
z = load [0x00]
```



Explicit Branches @ Resolution

Cause:

The resolution of a mis-speculation triggers a pipeline squash and alternation of control flow

```
if (secret) {
    y++;
}
z = load [0x00]
```

```
secret != prediction
```

- \rightarrow squash
- → load executes twice!





Covert channels

Explicit channels:

Secret inputs are directly leaked by operand-dependent hardware resource usage

Examples:

memory loads data-dependent arithmetic

Implicit channels:

Secret inputs are indirectly leaked by how (or that) one or several instructions execute

Explicit branches Implicit branches

Example:

Store-load pairs

Examples: Branch/jump instructions Leak on Leak on resolution





New!

Covert channels

Explicit channels:

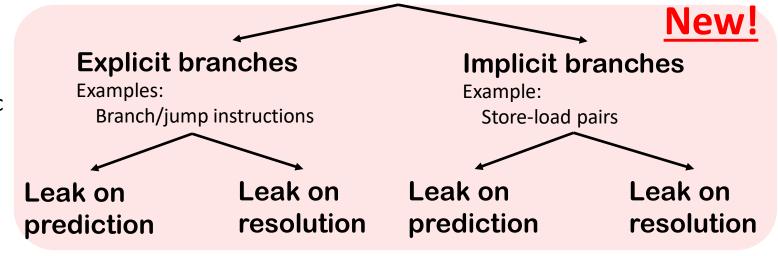
Secret inputs are directly leaked by operand-dependent hardware resource usage

Examples:

memory loads data-dependent arithmetic

Implicit channels:

Secret inputs are indirectly leaked by how (or that) one or several instructions execute







Implicit Branches

```
store [secret] = foo;
bar = load [0x00];
```



Implicit Branches

Cause:

Non-control flow instructions create branch-like behaviors.

```
store [secret] = foo;
bar = load [0x00];
```



Implicit Branches

Cause:

Non-control flow instructions create branch-like behaviors.

```
store [secret] = foo;
 bar = load [0x00];
               Can be thought as:
if (secret == 0x00) {
    forward from store queue
else
    cache load [0x00]
```



Basic idea: taint all the secrets

- Speculatively accessed data (secrets by definition)
- And their dependents



Basic idea: taint all the secrets

- Speculatively accessed data (secrets by definition)
- And their dependents

```
(addr < N) {
 // access instruction
 a = load [addr];
 // simple arithmetic
 b = a + 4;
 // cache/mem covert channel
 load [b];
```





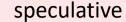
Basic idea: taint all the secrets

- Speculatively accessed data (secrets by definition)
- And their dependents

STT taints:

Output of speculative access instructions (a)

```
(addr < N) {
 // access instruction
 a = load [addr];
 // simple arithmetic
 b = a + 4;
 // cache/mem covert channel
 load [b];
```







Basic idea: taint all the secrets

- Speculatively accessed data (secrets by definition)
- And their dependents

STT *taints*:

- 1) Output of speculative access instructions (a)
- Output of instructions with tainted inputs (b)

```
(addr < N) {
 // access instruction
 a = load [addr];
 // simple arithmetic
 b = a + 4;
 // cache/mem covert channel
 load [b];
```





Basic idea: taint all the secrets

- Speculatively accessed data (secrets by definition)
- And their dependents

STT *taints*:

- 1) Output of speculative access instructions (a)
- Output of instructions with tainted inputs (b)

```
Resolved!
(addr < N) {
 // access instruction
 a = load [addr];
 // simple arithmetic
 b = a + 4;
 // cache/mem covert channel
 load [b];
```





Basic idea: taint all the secrets

- Speculatively accessed data (secrets by definition)
- And their dependents

STT *taints*:

- Output of speculative access instructions (a)
- Output of instructions with tainted inputs (b)

STT **untaints** when:

1) A speculative access instruction becomes nonspeculative (a)

```
Resolved!
if (addr < N) {
    // access instruction
    a = load [addr];
    // simple arithmetic
    b = a + 4;
    // cache/mem covert channel
    load [b];
```





Basic idea: taint all the secrets

- Speculatively accessed data (secrets by definition)
- And their dependents

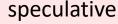
STT taints:

- Output of speculative access instructions (a)
- Output of instructions with tainted inputs (b)

STT **untaints** when:

- A speculative access instruction becomes nonspeculative (a)
- An instruction has all its input untainted (b)

```
Resolved!
if (addr < N) {
    // access instruction
    a = load [addr];
    // simple arithmetic
    b = a + 4;
    // cache/mem covert channel
    load [b];
```







Basic idea: taint all the secrets

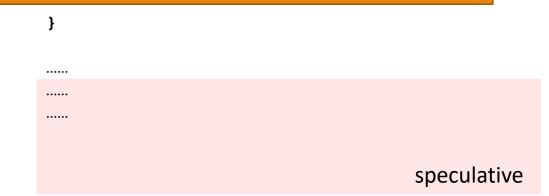
Speculatively accessed data (secrets by definition)

if (addr < N) { ← Resolved!

Data is tainted \rightarrow Data is speculative (not necessarily other way around)

STT **untaints** when:

- A speculative access instruction becomes nonspeculative (a)
- 2) An instruction has all its input untainted (b)







Microarchitect Identifies ...

Instructions forming explicit channels

• E.g. load, data-dependent arithmetic

Instructions forming implicit channels

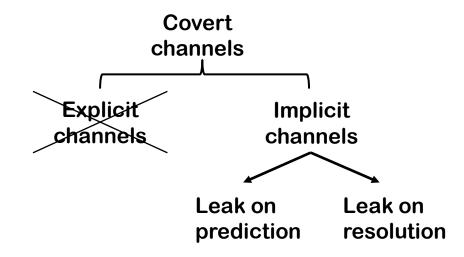
E.g. control-flow instructions, store-load pairs



Blocking Covert Channels

Explicit channels:

Delay execution until operands untainted (e.g., load address)





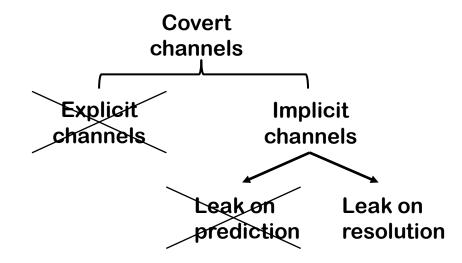
Blocking Covert Channels

Explicit channels:

Delay execution until operands untainted (e.g., load address)

Implicit channels:

Delay predictor update until branch predicate untainted





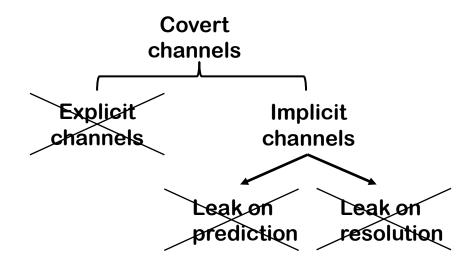
Blocking Covert Channels

Explicit channels:

Delay execution until operands untainted (e.g., load address)

Implicit channels:

- Delay predictor update until branch predicate untainted
- Delay resolution until branch predicate untainted





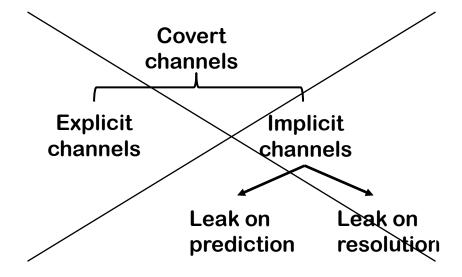
Blocking Covert Channels

Explicit channels:

Delay execution until operands untainted (e.g., load address)

Implicit channels:

- Delay predictor update until branch predicate untainted
- Delay resolution until branch predicate untainted





What speculative work can we safely do?

- 4
- Safe to execute all instructions w/ untainted operands
- Safe to execute safe (no explicit channel) instructions w/ tainted operands
- Safe to predict on implicit/explicit branches w/ tainted predicates Note: predictors have high accuracy.

```
a = 0
if (secret) a+=CACHE_LN_SZ
load(a) // covert channel
```

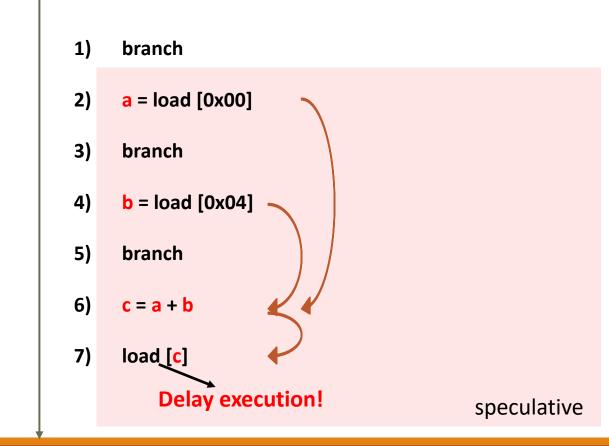
PC = non-sensitive

- → Predictor state = non-sensitive
- → Safe to predict on branch ©

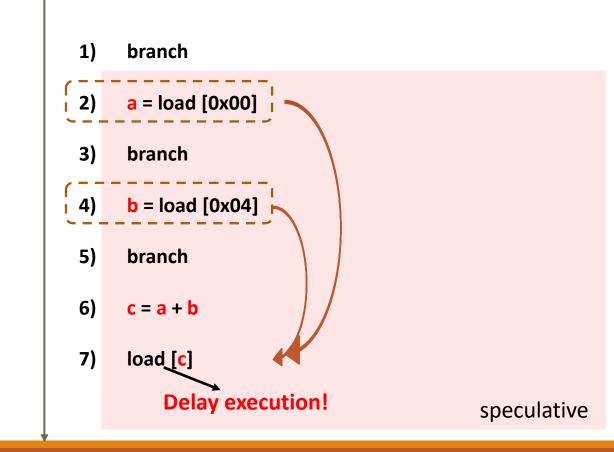
Hardware Implementation of STT



program order

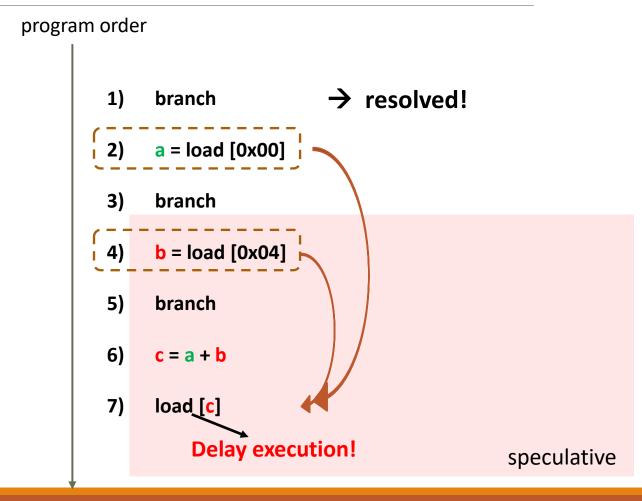


program order

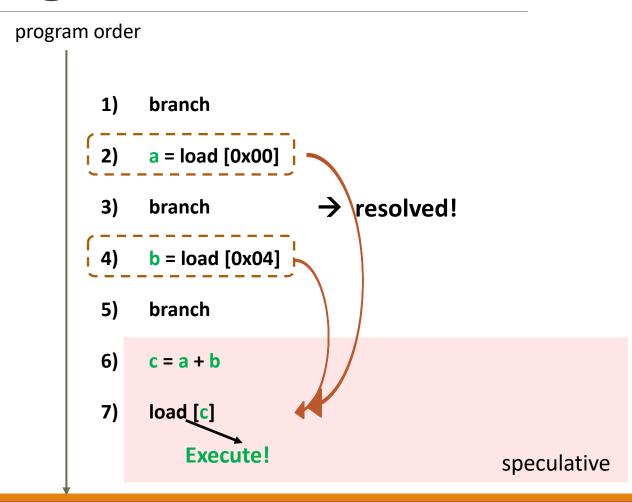


Observation: All instructions turn nonspeculative in-order program order branch a = load [0x00]branch b = load [0x04]branch c = a + b7) load[c] **Delay execution!** speculative

Observation: All instructions turn nonspeculative in-order



Observation: All instructions turn nonspeculative in-order





Observation: All instructions turn nonspeculative in-order

Each instruction tracks the "youngest access instruction" it depends on -- "Youngest Root of Taint" (YRoT)

program order

- 1) branch
- 2) a = load [0x00]
- 3) branch
- 4) b = load [0x04]
 - 5) branch
 - 6) c = a + b

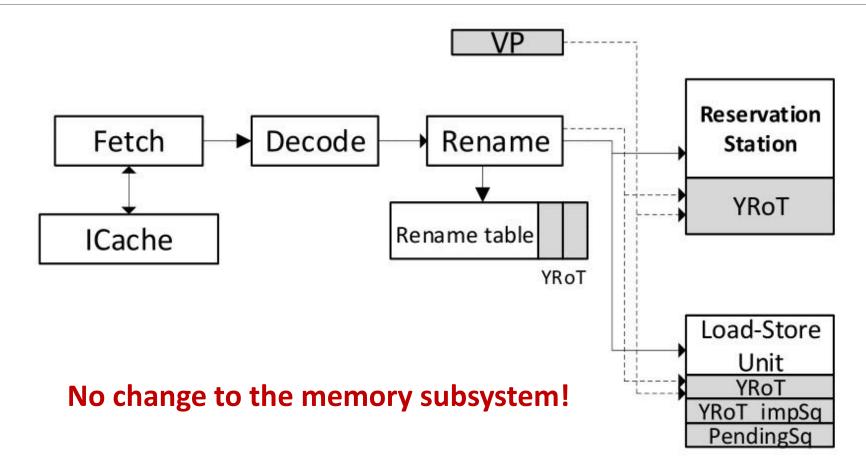
7) load [c]

Execute!

YRoT of 7 is 4

speculative





Security Evaluation

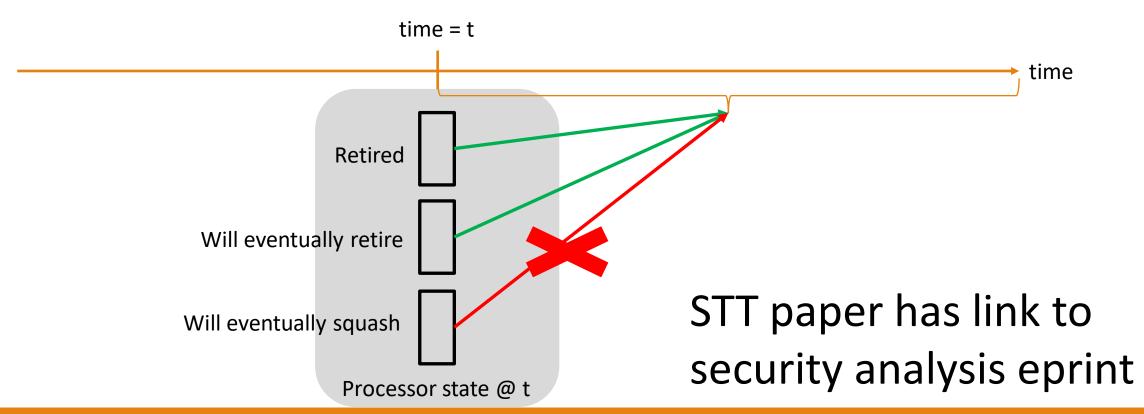
Security definition:

Arbitrary speculative execution can only leak retired register file state (not arbitrary program memory)

No read gadgets!

Security Evaluation

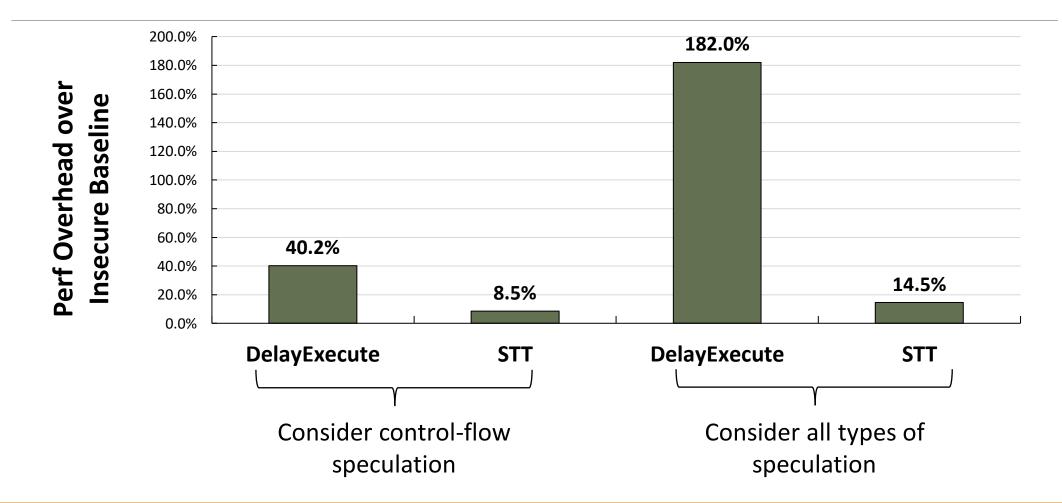
STT enforces a non-interference property w.r.t speculatively accessed data:







Performance Evaluation on SPEC2006





Summary

STT Blocks leakage of speculatively accessed data over any uarch covert channels with:

- 1) High performance
- 2) Provable security protection
- 3) No software change; No memory subsystem change



Part 2: Speculative Data-Oblivious Execution (SDO)

PERFORMANCE OPTIMIZATION FRAMEWORK FOR STT





Where does overhead come from in STT?

Explicit channels (a.k.a. transmit instructions):

Delay execution until operands untainted (e.g., load address)

>90% of overhead

Implicit channels:

- Delay predictor update until branch predicate untainted
- Delay resolution until branch predicate untainted

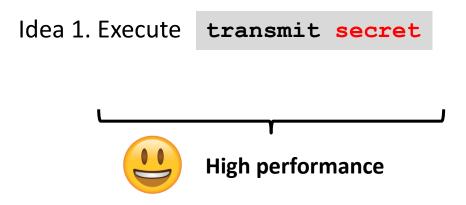
```
if (addr < N) { // speculation

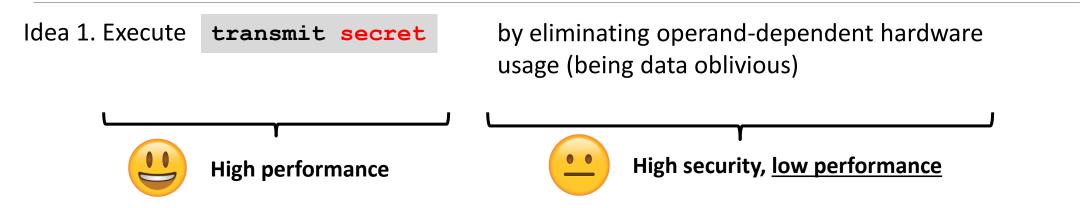
    // access instruction
    secret = load [addr];

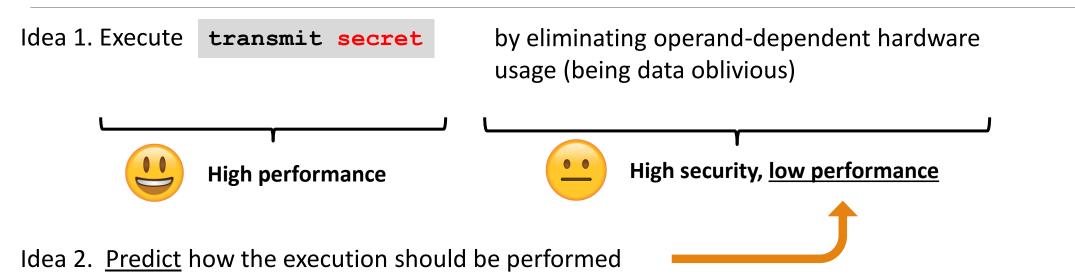
    // transmit instruction
    transmit secret;
}

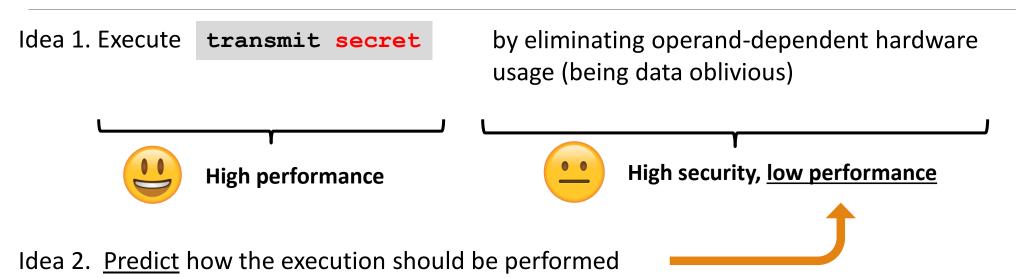
E.g., loads, floating point, ...</pre>
```

→ Delay execution







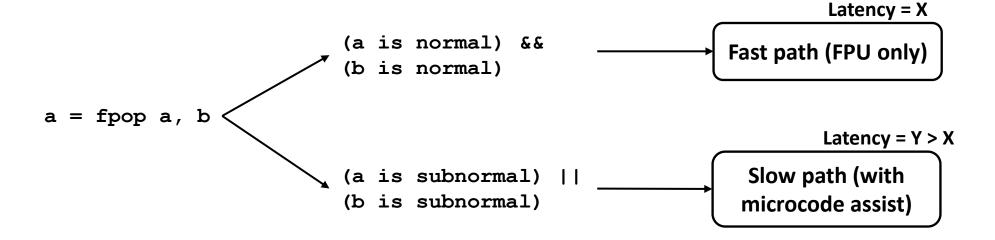


Problem: combining idea 1 & 2 creates security problems

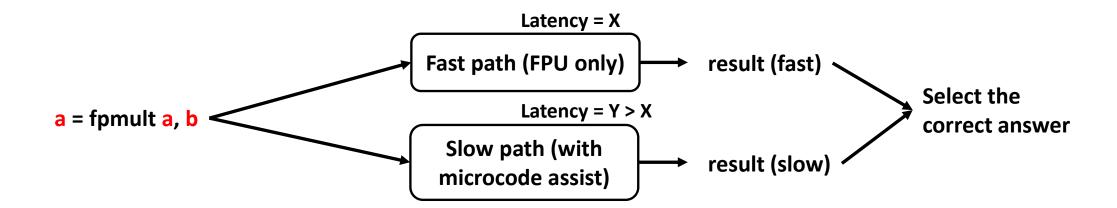
Solution: build on top of Speculative Taint Tracking (STT)

Example: Subnormal Floating-point Operation

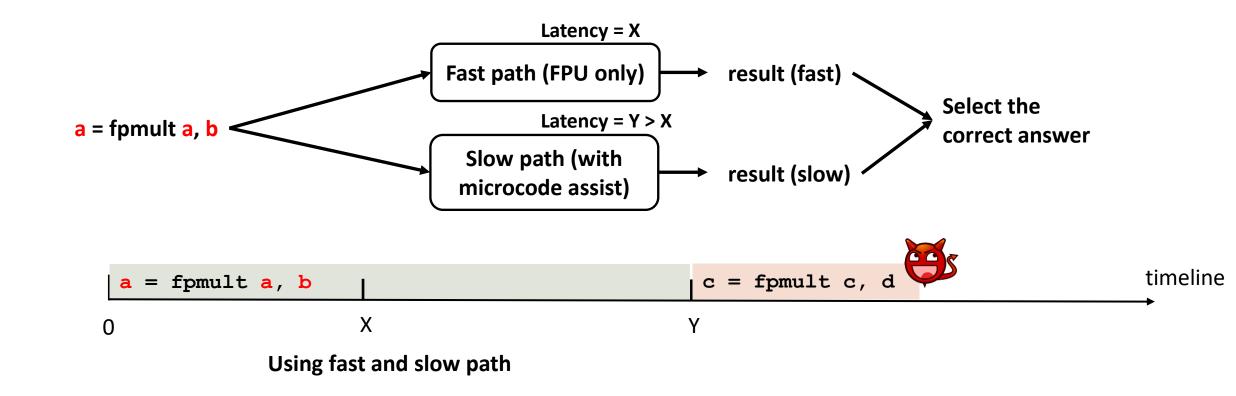
- Double-precision floating point
 - Normal input: (2.23e-308, 1.79e308), processed by Floating-Point Unit (FPU)
 - Subnormal input: (4.9e-324, 2.23e-308), requiring microcode assist



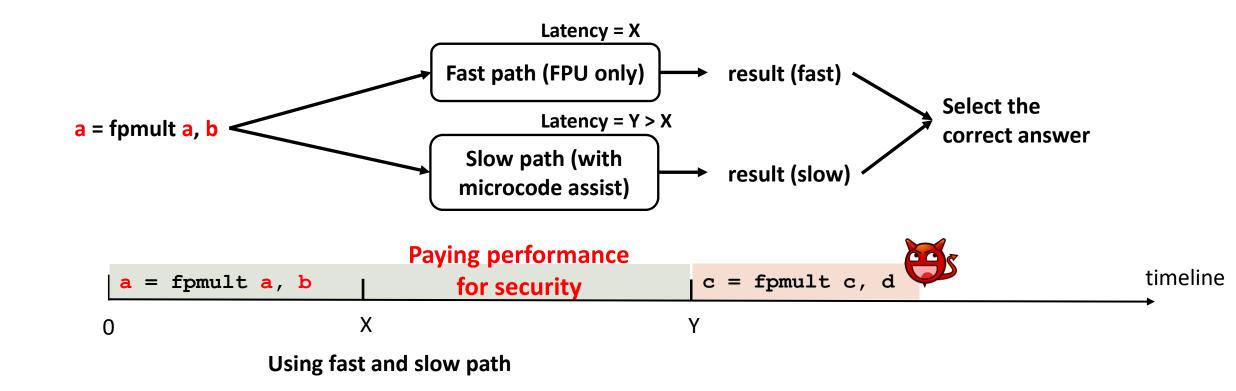
Idea 1: Being Data Oblivious

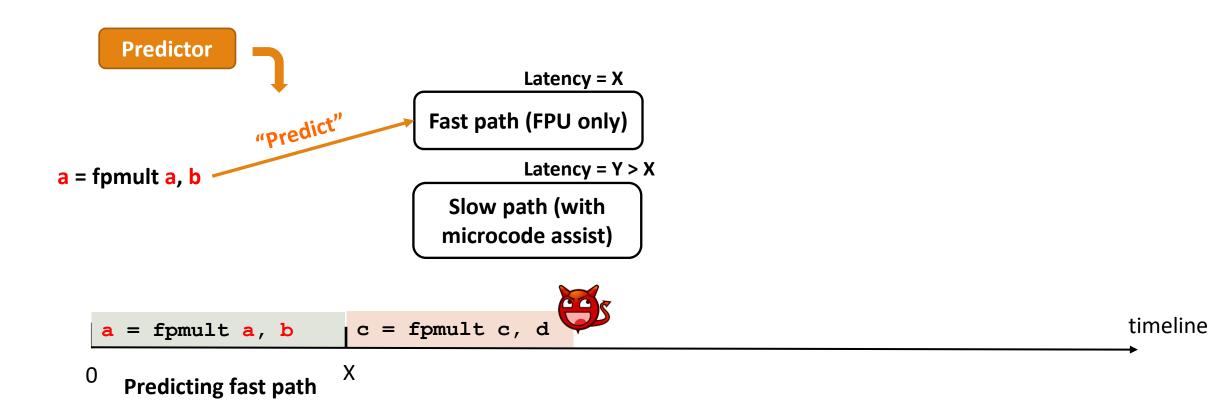


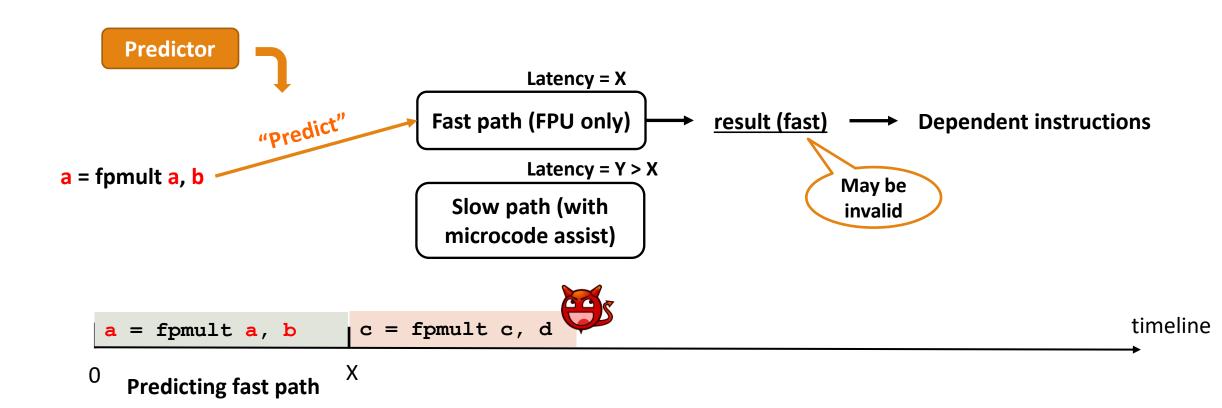
Idea 1: Being Data Oblivious

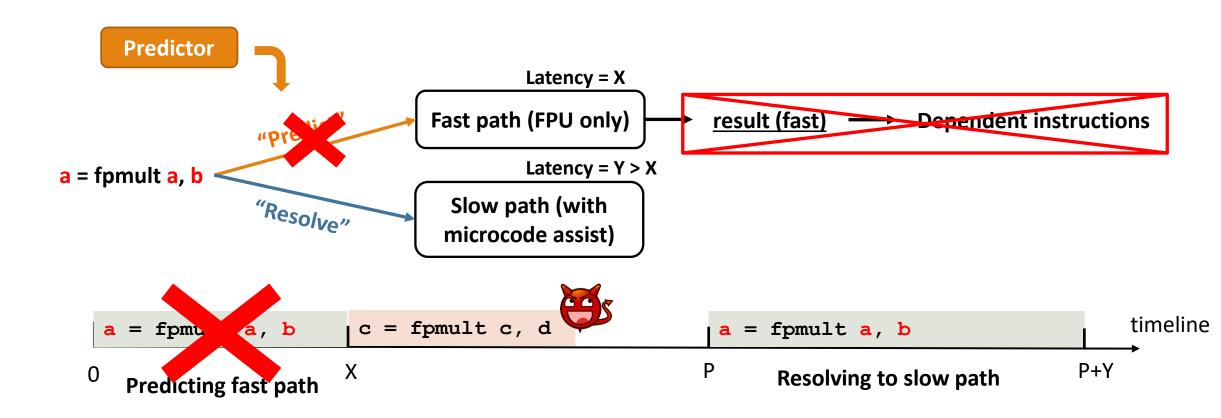


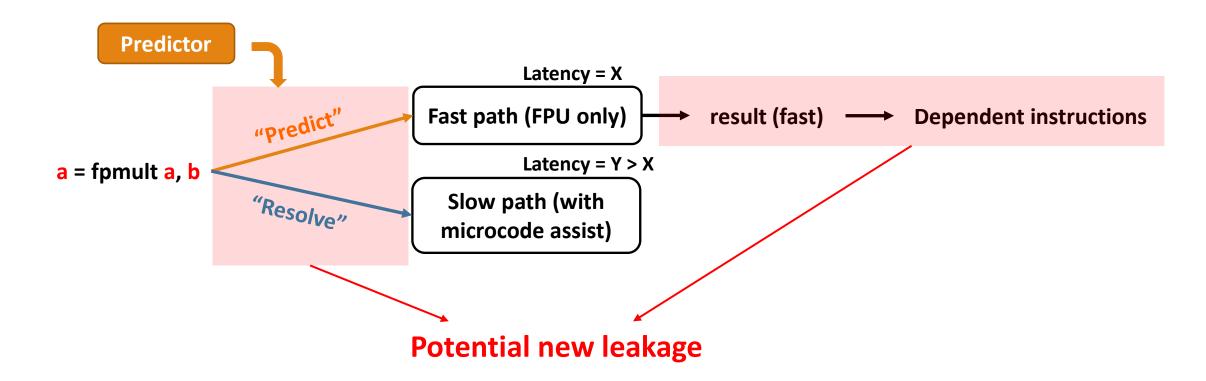
Idea 1: Being Data Oblivious

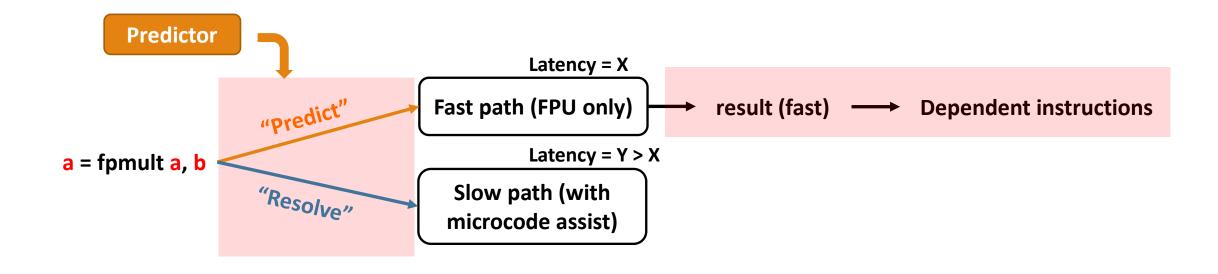




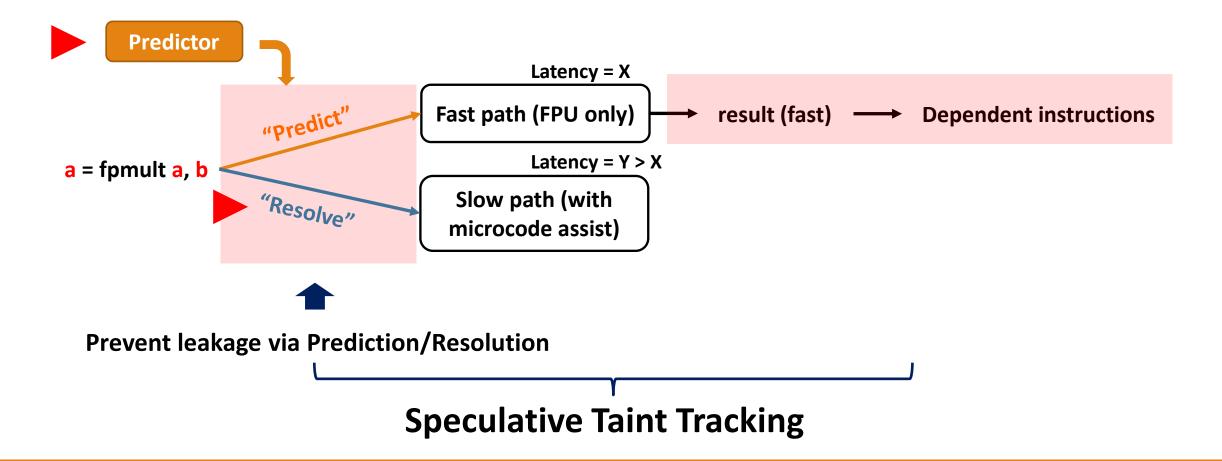


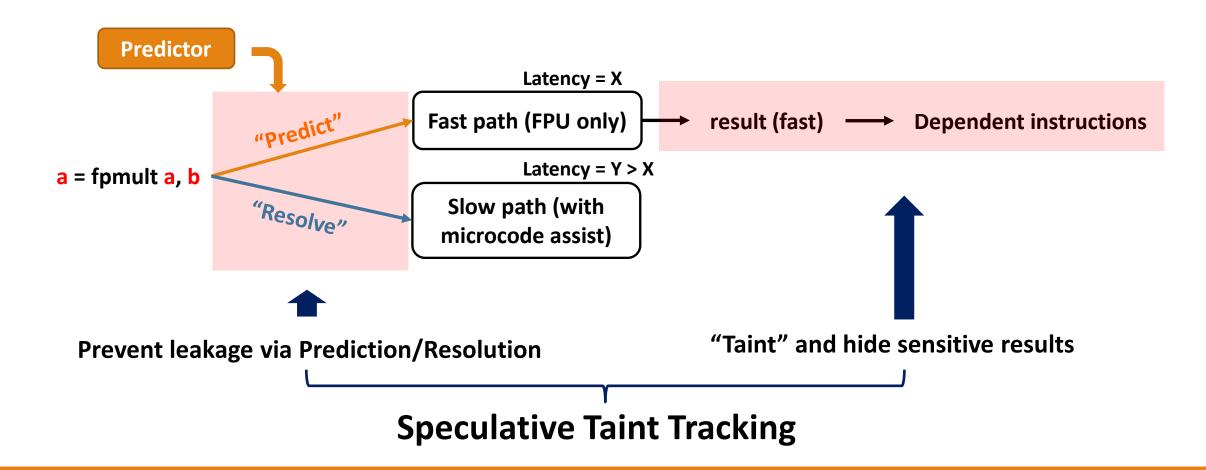






Speculative Taint Tracking





How STT "prevents leakage via prediction/resolution":

- Never update predictors with any secret information
- Delay resolution until safe

How STT "prevents leakage via prediction/resolution":

- Never update predictors with any secret information
- Delay resolution until safe

How STT "taints and hides sensitive results":

- Sensitive data is marked tainted
- Taint propagates through program dataflow
- Transmitters with tainted arguments are handled safely

Applying STT for Security

How STT "prevents leakage via prediction/resolution"



STT Makes Prediction SAFE Again!



We build predictors to reduce defense overhead

- Tailit propagates tillough program uatanow
- Transmitters with tainted arguments are handled safely

Speculative Data Oblivious Execution (SDO)

Idea 1. Safely execute transmitters in a data-oblivious (DO) manner

Idea 2. Predict how the execution should be performed



<u>Data Oblivious variants</u> + <u>Predicting which variant</u>

+ Safe Prediction with STT

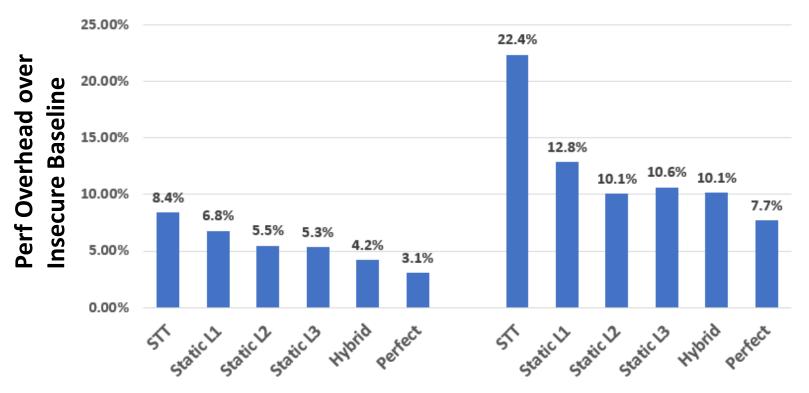
=

SDO



Net result: execute unsafe transmitters early and safely

Performance Evaluation on SPEC2017



"Spectre" attack model Consider control-flow speculation

"Futuristic" attack model Consider all types of speculation

Transmitters:

- Load
- Floating-point multiplication
- Floating-point division

Static L1: always predicting DO-ld_{L1}
Static L2: always predicting DO-ld_{L2}
Static L3: always predicting DO-ld_{L3}
Hybrid: using the hybrid predictor

Perfect: prediction is accurate and precise

Conclusion

<u>Data Oblivious variants</u> + <u>Predicting which variant</u> + <u>Safe Prediction with STT</u>

Safe, early execution of transmitters

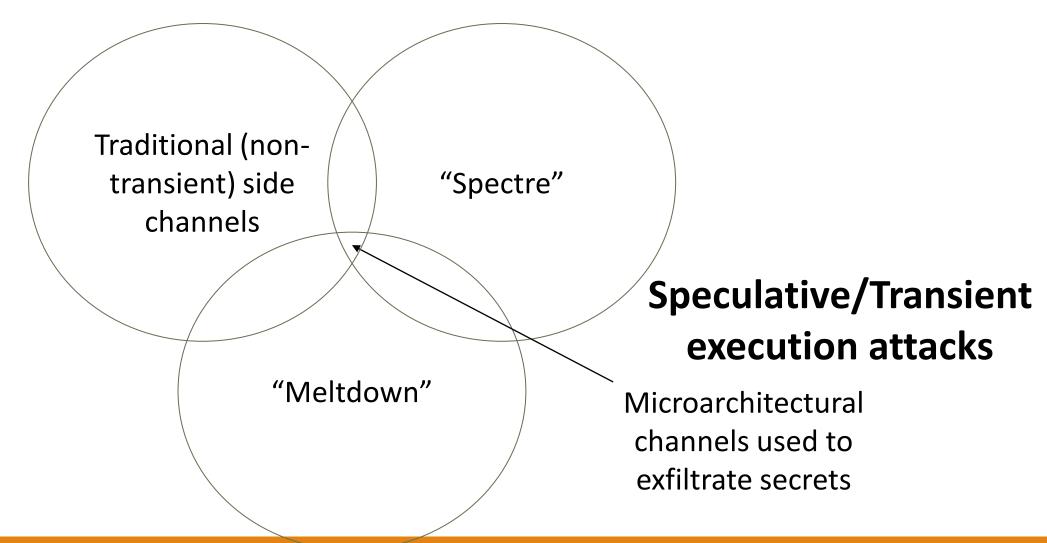
Part 3: Where things are going (my view) + some new read gadgets



Pre 2018

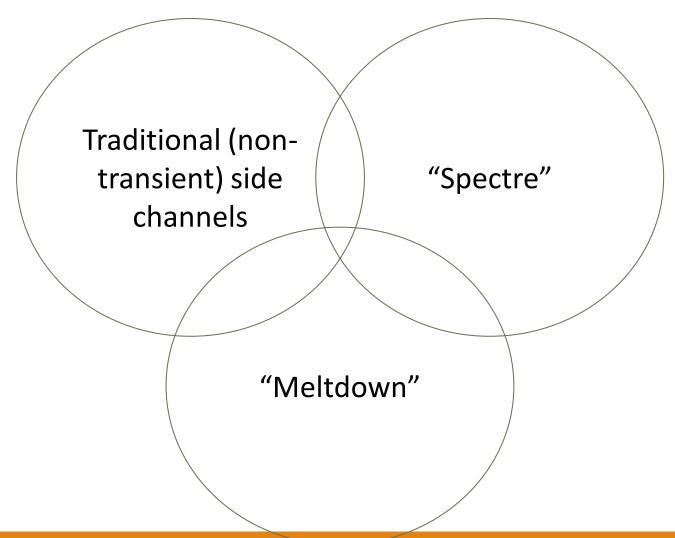
Traditional (nontransient) side channels



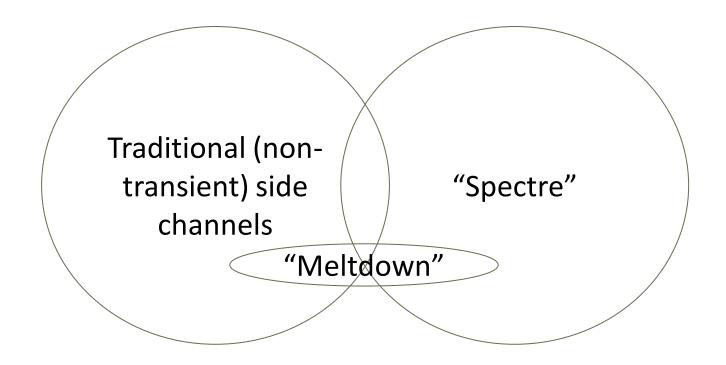




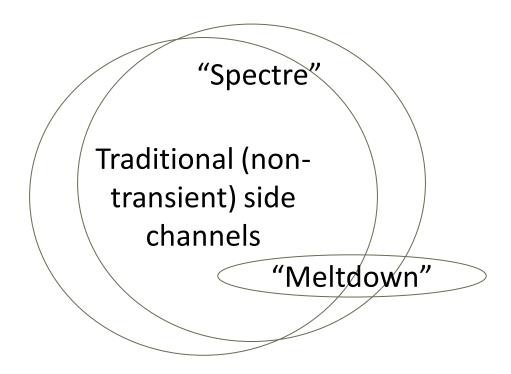




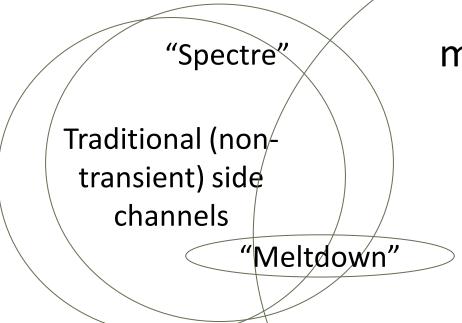












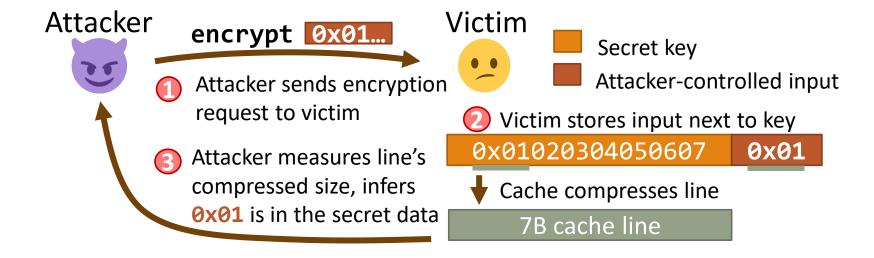
How does other microarchitecture leak privacy?



RAMBleed (RowHammer)

Safecracker (Compressed caches) [ASPLOS'20]

Compressed Cache Attacks



Read Gadgets from Compressed Cache

Co-locate attacker data w/ secret data \rightarrow leak secret data Numerous ways to co-locate data.

```
HEARTBLEED-LIKE

p = malloc(SZ);
memcpy(p, usr_data, SZ);
```

BROP-LIKE

Given:

- * re-startable service
- * "buffer overflow"
- 1.) Overflow buffer to size N, guess byte N+1
- 2.) Repeat (1) until byte N+1 leaked
- 3.) N++; Goto (1).





Conclusion

In crypto/info flow, we usually ask: do *any* secret bits leak? In HW, need to ask when *all* bits can leak.

Need new abstractions/defenses to reason about leakage.

