Protecting Users' Privacy when Tracing Network Traffic

Stefan Saroiu and Troy Ronda
University of Toronto

Challenge: Protect Users' Privacy

- Network tracing today must capture payloads:
  - Challenge: protect users' privacy

- Typically, privacy protected via 3-step process:
  1. Gather raw data,
  2. Anonymize it off-line by hashing information
     - Preserve some info: IP prefix-sharing, object sizes, etc..
  3. Throw-away raw data

- Trace analysis is done on anonymized data
  - Anonymized data could become publicly available

3-step process is inadequate from a privacy standpoint!

3-Step Anonymization Doesn't Work

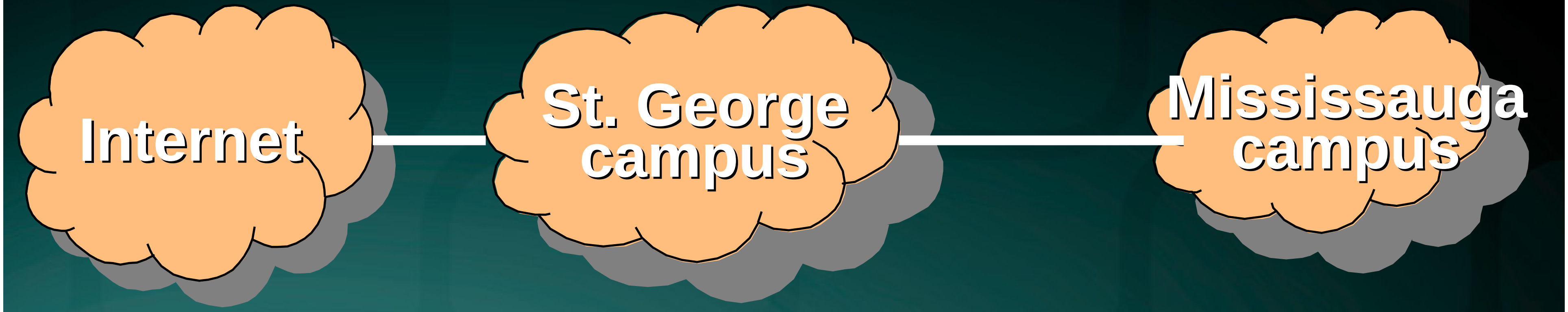
- Known mapping attacks:
  - e.g., one IP address shares no prefix with all others
  - e.g., CEO is biggest recipient of e-mail
- Inferred mapping attacks:
  - e.g., we could guess what websites are top 10 most popular
    - google.com, www.utoronto.ca, etc..
  - e.g., one 700MB file became a hot download on 11/3/2006
    - The Borat movie was released on the same date
- Data injection attacks:
  - Attacker injects carefully constructed traffic
  - Traffic easy to distinguish in hashed trace
- Crypto attacks:
  - Finding MD5 collisions takes 8 hours on a laptop today!!!
  - Would today hashed trace be trivial to break 20 years from now?

Even More Attacks are Possible

- Attacks on tracing infrastructure:
  - Network intrusions
  - Physical intrusions
- Unanticipated attacks:
  - Hard to foresee future ways to attack anonymization scheme
  - e.g., OS could be revealed based on ACKs' timestamps
- Legal complications (attacks?):
  - Tracing infrastructure could be subpoena-ed
  - Precedents exist: e.g., RIAA vs. Verizon

Lessons Learned

- No plaintext data can be written to disk. Ever.
  - Subpoenas can reveal whole profiles
    - Very serious attack with serious privacy implications
- Gathered traces cannot be made public
  - Mapping attacks could reveal private information
  - Subject to future crypto attacks
    - a PDA will break MD5 in under 1 second in 20 years
  - Unanticipated attacks are problematic

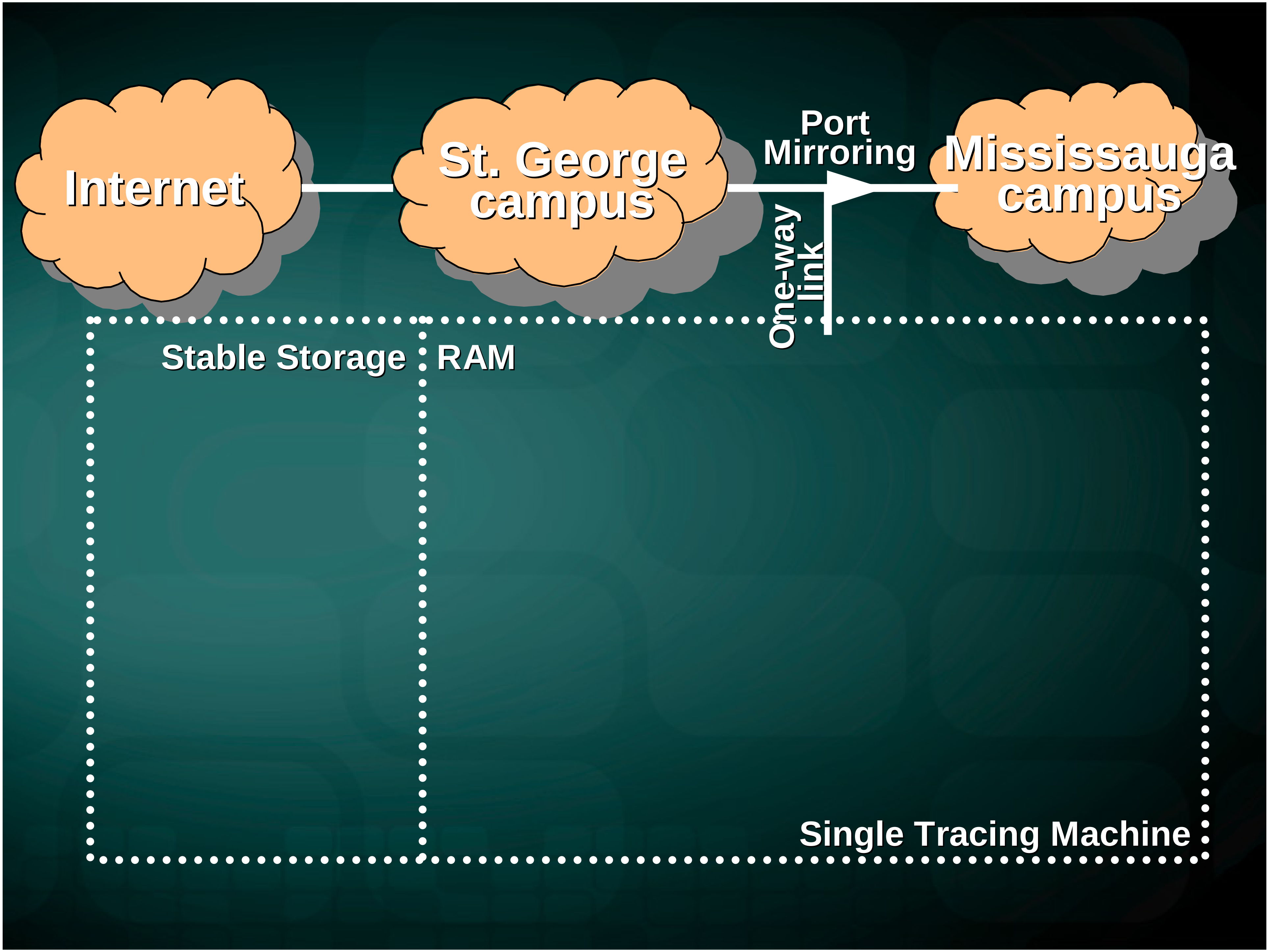


Internet

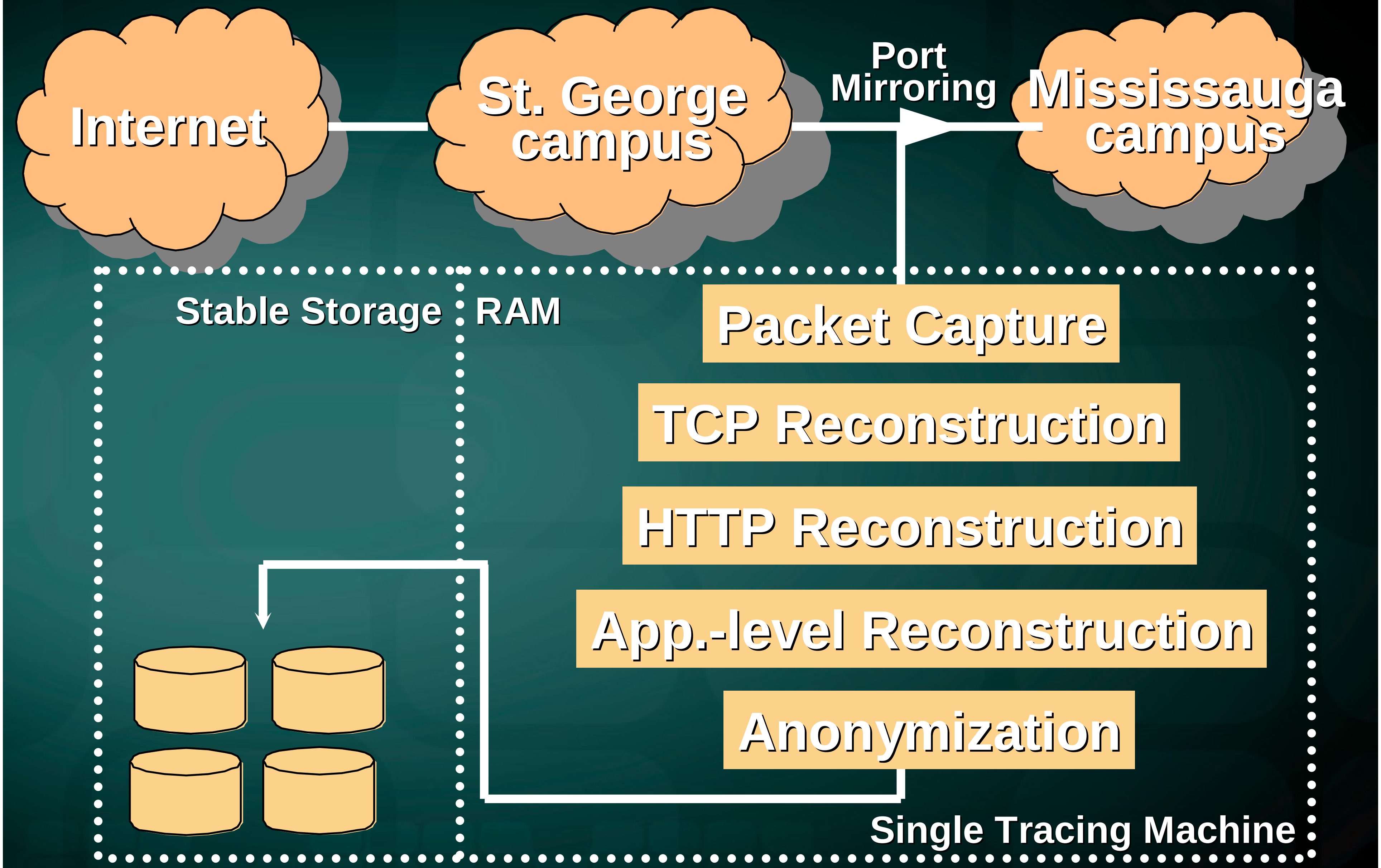
St. George campus

Mississauga campus

Picture not up to scale!



Internet
St. George campus
Port Mirroring
Mississauga campus
One-way link
Stable Storage
RAM
Single Tracing Machine



Internet

St. George campus

Port Mirroring

Mississauga campus

Stable Storage

RAM

Packet Capture

TCP Reconstruction

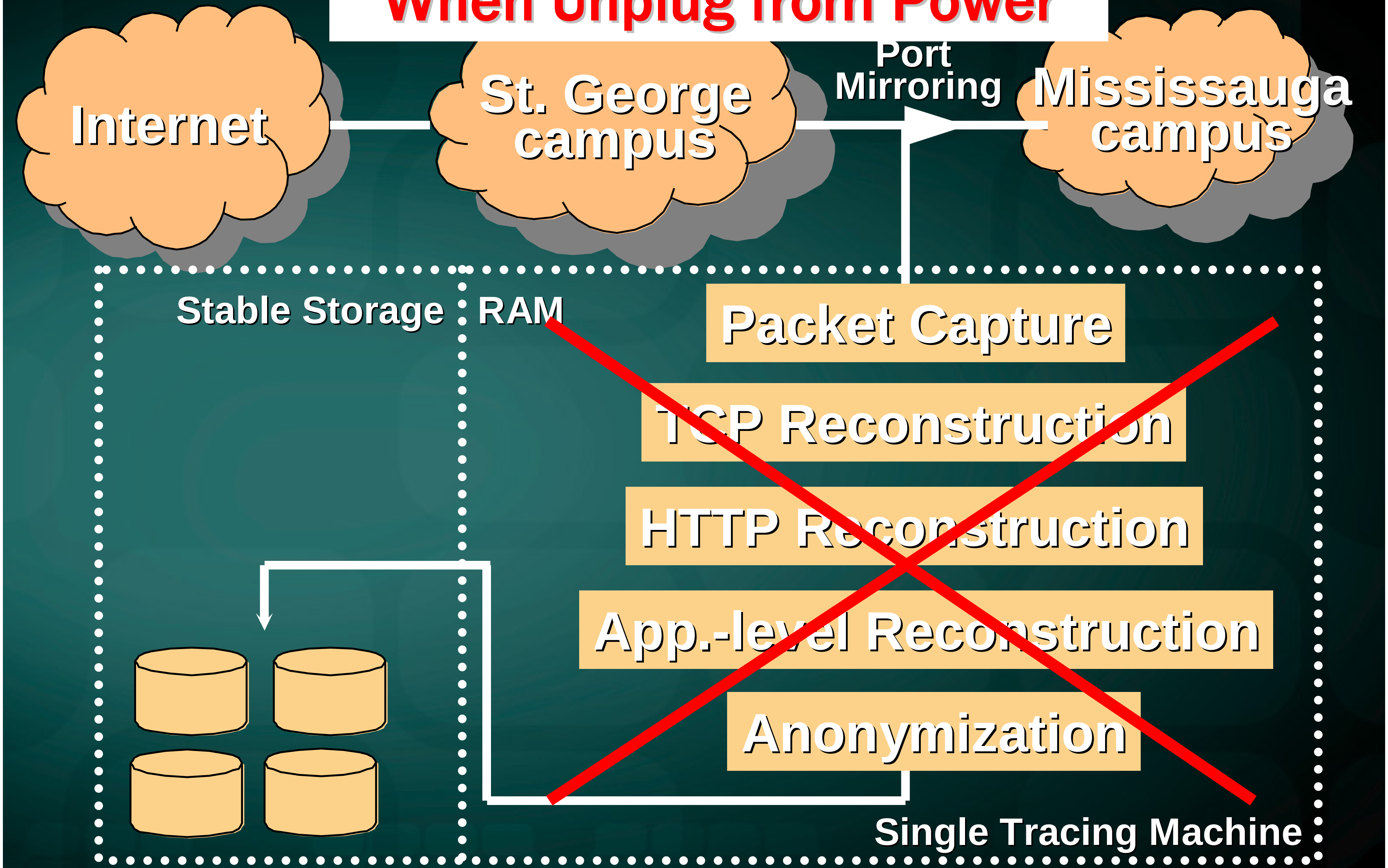
HTTP Reconstruction

App.-level Reconstruction

Anonymization

Single Tracing Machine

When Unplug from Power



Internet

St. George campus

Port Mirroring

Mississauga campus

Stable Storage

RAM

Packet Capture

TCP Reconstruction

HTTP Reconstruction

App.-level Reconstruction

Anonymization

Single Tracing Machine

Summary

- Our infrastructure protects against:
  - Intrusion attacks
    - Disconnected from Internet
  - Legal attacks to recover raw data
    - All raw data manipulation done in RAM
  - Mapping, crypto, unanticipated, data injection attacks
    - Traces will not be made publicly available

- Mapping, crypto, unanticipated attacks still possible if anonymized trace is subpoena-ed
  - Once analysis complete, destroy trace permanently

Phishing Measurement Statistics (Very Preliminary)

- Tracing 200Mbps and approximately 5K users
  - 20GB of data collected per day
- Longest uninterrupted trace: 56 hours
- E-mail usage statistics (spam)
  - 213 Hotmail users, 721 messages received
  - 22 (3%) spam in Inbox (missed by Hotmail's filters)