Computer Science Techniques for Fintech: Recent research introduction

Wei Xu

Institute for Interdisciplinary Information Sciences

Tsinghua University

About Me - Wei Xu 徐葳

- Tsinghua & Penn (B.S) 1999-2003
- Berkeley (M.S. and Ph.D.) 2003 - 2010
  - Advisors: David Patterson and Armando Fox
- Google 2010 - 2013
- Joined Tsinghua in 2013
  - Assist. Prof. and Assist. Dean @ Institute for Interdisciplinary Information Sciences (IIS)
- Research Area
  - Distributed systems + Machine learning
 - Interdisciplinary "Big data" Applications, esp. Fintech

Smart data center networking and operations
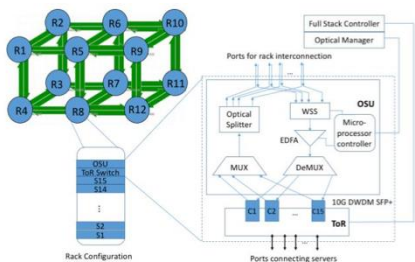
(a) DFabric architecture, rack configuration and OSU structure.
OSU: optical switching unit; ToR switch: top-of-rack switch; MUX: multiplexer, DeMUX: demultiplexer, WSS: wavelength selective switch; EDFA: Erbium-doped fiber amplifier.
S1-S15: servers; C1-C15: 10G DWDM SFP+ transceivers

(b) Photo of DFabric
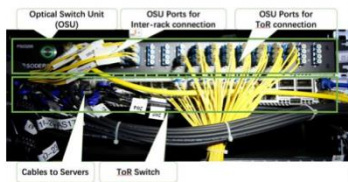
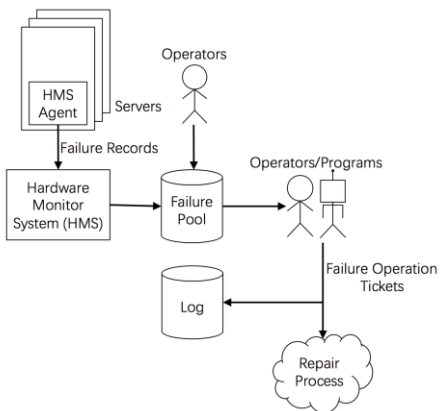(c) Photo of OSU connected with ToR switch
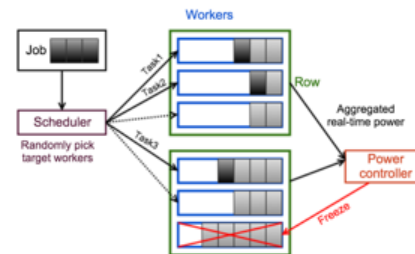
Optical data center optimization to improve tail latency [Sigcomm'16, OFC'16]

Data center power prediction and scheduling, improving computing capacity by 17% [EuroSys16]

Large scale data center hardware failure analysis [DSN 17 BEST PAPER AWARD]

Automatic system operation knowledge discovery and natural language-based Q&A [APsys16-BEST PAPER]

"Fintech" is about different things in China vs. in the US

- In the US: more efficiency
  - Tax filing
  - Compliance

- In China: All about new business models
  - "Internet finance"
  - Credit scores from alternative data
  - Mobile payment
  - e-commerce + payment + loans
  - Supply chain financing

Differences in Fintech development

US: navigating through the mature market

China: flying into the unknown, fast

Significant challenges in the infrastructure of Fintech in China

Problems：

- Trust
- Risk management
- Privacy
- Regulations
- Data monopoly

Consequences：

"Policy Risk"

aka. Blaming the government

Key techniques useful in Fintech

- AI and big data

- Distributed systems and consensus

- Crypto techniques

Project 1：Data-driven anti-fraud

- Challenge
  - No centralized credit bureau
  - Black markets provides fake entities, fake phones and AI techniques
  - Rendering the rule / blacklist based approaches obsolete
  - Similar problems in anti-money laundering
- Our solution
  - Unsupervised / semi-supervised fraud detection
  - Detecting group based fraud behaviors

Fraud example：Group registration

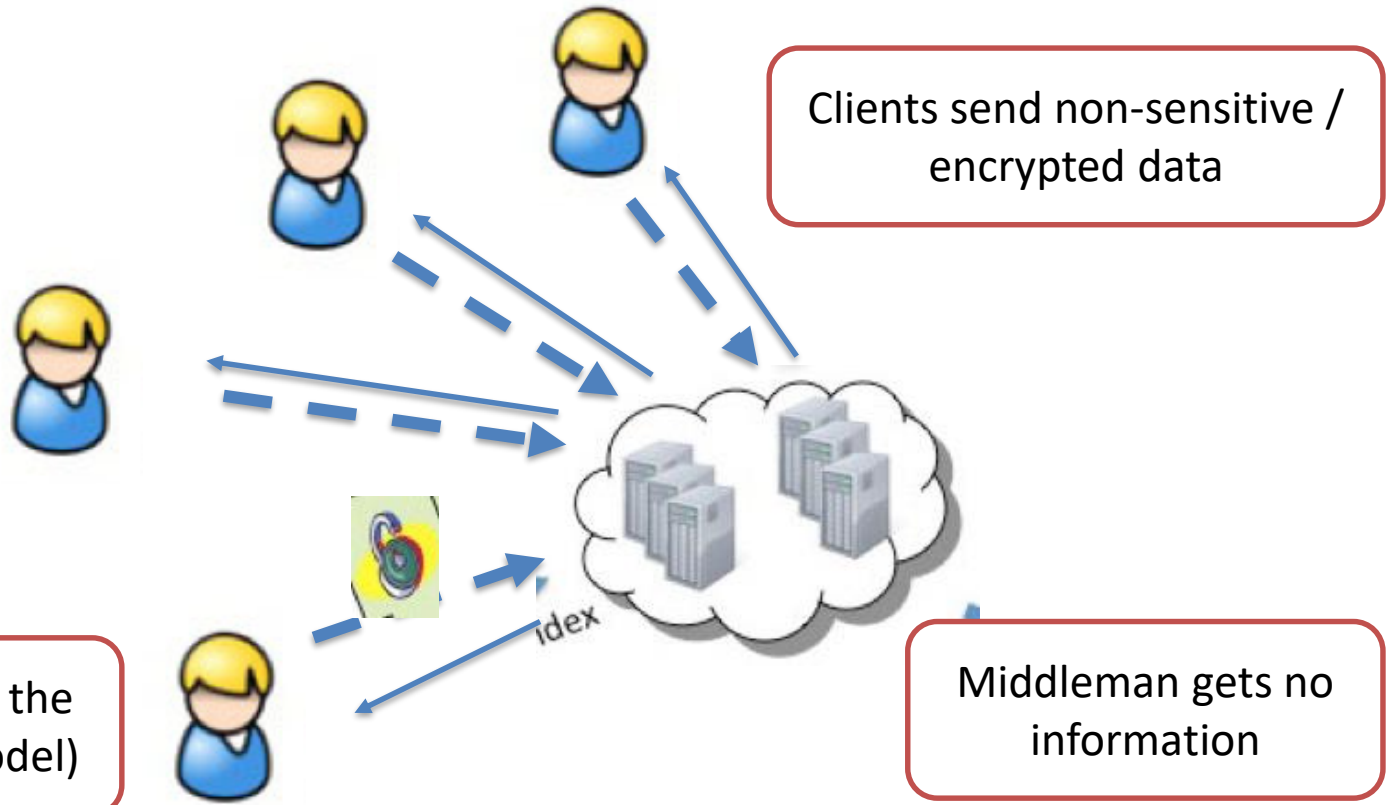| 注册城市	IP子网	操作系统	手机前缀	手机所在地	注册来源	注册时间长度	注册时间
大连市	106.3.88.0		1589994****	东莞	1	13159	2017-04-18 20:00:00
宜春市	220.176.164.0		1589994****	东莞	1	15501	2017-04-18 19:00:00
咸阳市彬县	117.23.44.0		1589994****	东莞	1	18715	2017-04-18 19:00:00
宁波市	125.115.72.0		1589994****	东莞	1	26079	2017-04-18 19:00:00
南充市	112.194.68.0		1589994****	东莞	1	23050	2017-04-18 18:00:00
宣城市	183.165.72.0		1589994****	东莞	1	18155	2017-04-18 18:00:00
安庆市	112.122.72.0		1589994****	东莞	1	84591	2017-04-18 18:00:00
泸州市	124.161.112.0		1589994****	东莞	1	20692	2017-04-18 18:00:00
重庆	27.11.140.0		1589994****	东莞	1	13010	2017-04-18 18:00:00
泉州市	110.81.156.0		1589994****	东莞	1	20518	2017-04-18 18:00:00
德阳市	175.154.204.0		1589994****	东莞	1	18993	2017-04-18 17:00:00
赣州市	106.5.120.0		1589994****	东莞	1	94006	2017-04-18 17:00:00
成都市双流县	125.71.64.0		1589994****	东莞	1	14941	2017-04-18 17:00:00
达州市	175.153.140.0		1589994****	东莞	1	19511	2017-04-18 17:00:00
德阳市	175.154.204.0		1589994****	东莞	1	10874	2017-04-18 16:00:00
福州市	220.160.8.0		1589994****	东莞	1	14595	2017-04-18 16:00:00

Project 2：Privacy preserving multi-party computation

- People start to care about privacy, and more privacy-related regulations
- However, data exchange is often key to businesses
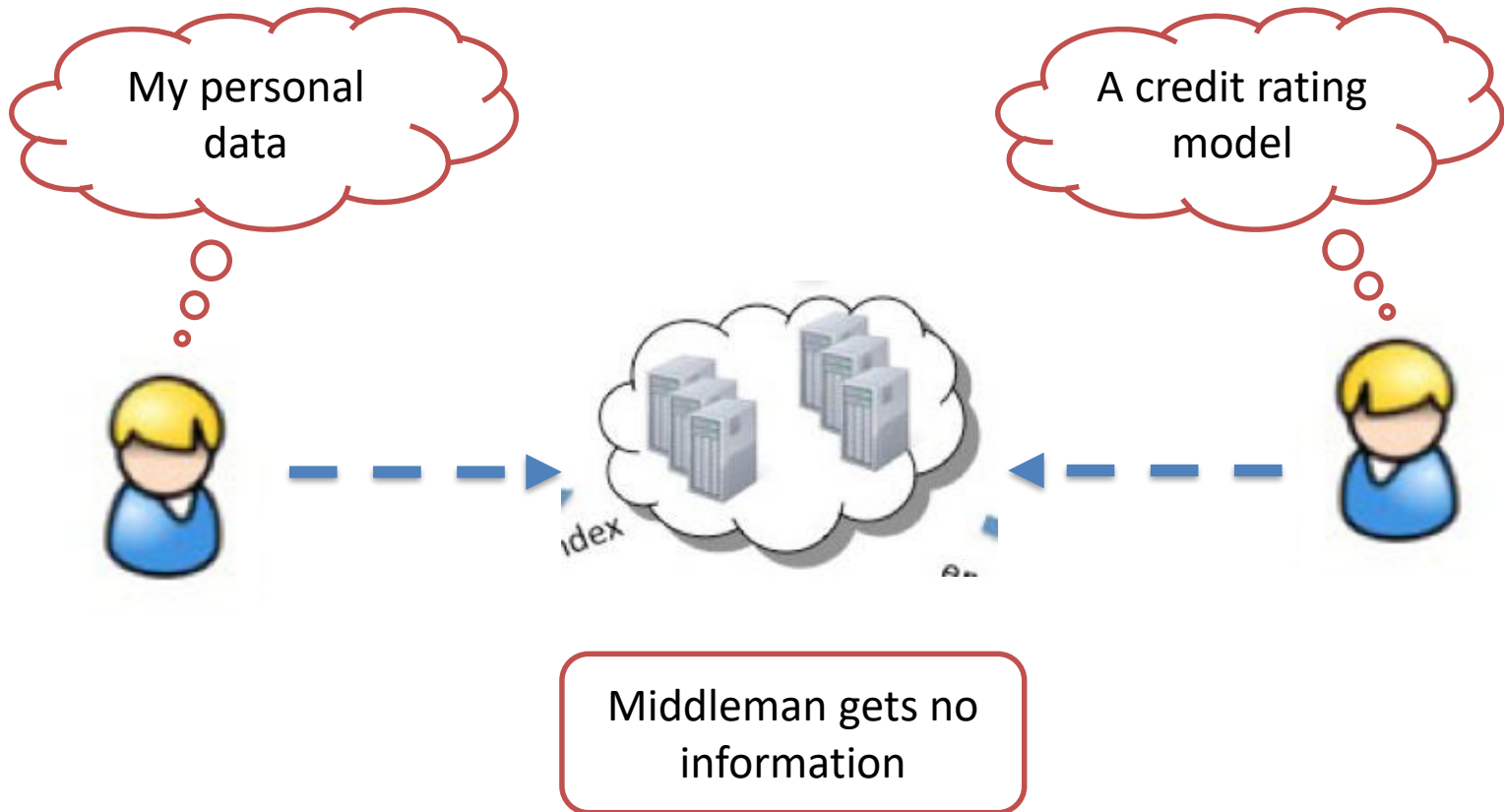- How can we solve the problem?

Scenario 1: Data market

Clients send non-sensitive / encrypted data

Participants get the result (e.g. a model)
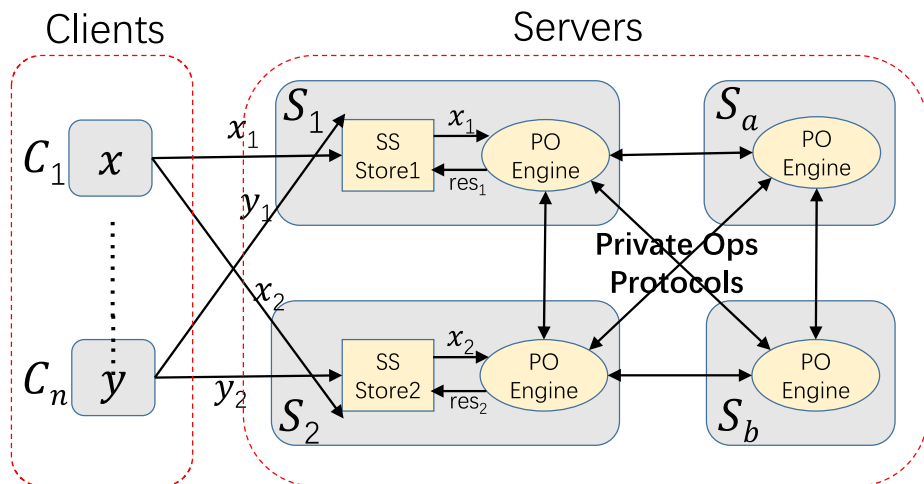
Middleman gets no information

Scenario 2：Computation on private model and data

Project 2：Privacy preserving multi-party computation

Clients

Servers

Scalable, easy to program secure multi party computation engine

Challenges

- Integrating multiple crypto tools and optimize for real system performance
- Easy Python interface, so we can port applications
- Efficient and Scalable

Real world performance in a deep learning application

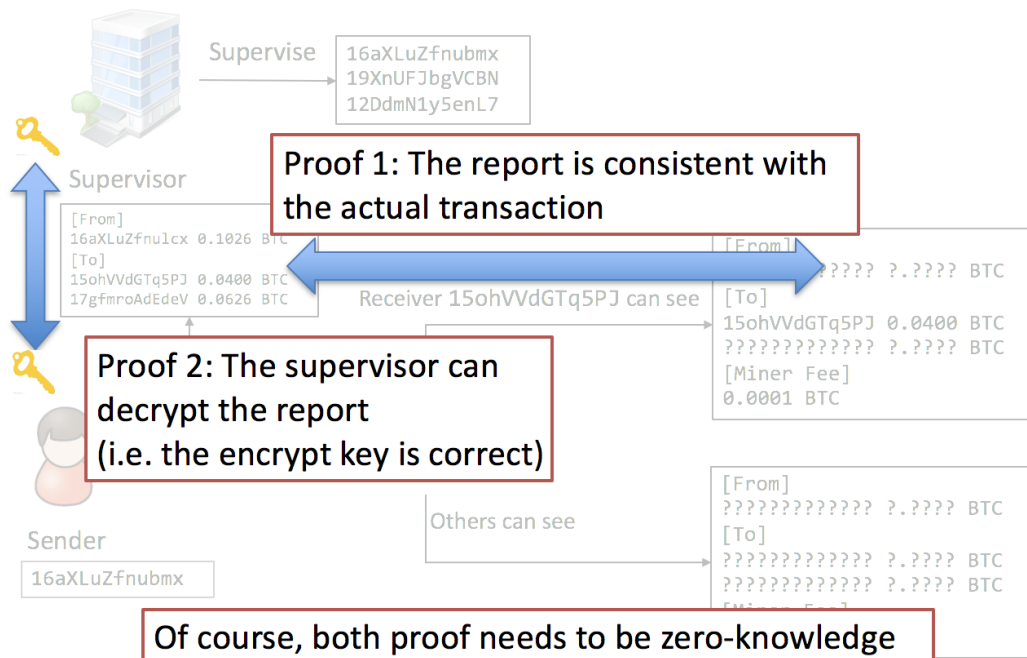| batch size	Feedforward NN	CNN	CNN + BN
1	1.48	1.57	1.58
10	0.34	1.18	1.18
100	0.24	1.15	1.16
1000	0.23	1.12	1.13
all	0.21	1.10	1.12

Project 3：Privacy and regulations, how to strike a balance?

- Regulators needs to see something
- But not everything

- E.g. Custodian bank and traders

- Can we ensure that?

Project 3：Privacy and regulations

Supervise
```
16aXLuZfnubmx
19XnUFJbgVCBN
12DdmN1y5enL7
```

Supervisor
```
[From]
16aXLuZfnu1cx  0.1026 BTC
[To]
15ohVVdGTq5PJ  0.0400 BTC
17gfmroAdEdeV  0.0626 BTC
```

Proof 1: The report is consistent with the actual transaction

Receiver 15ohVVdGTq5PJ can see

```
[From]
????? ?.???? BTC
[To]
15ohVVdGTq5PJ  0.0400 BTC
????????????? ?.???? BTC
[Miner Fee]
0.0001 BTC
```

**Proof 2: The supervisor can decrypt the report
(i.e. the encrypt key is correct)**

Others can see
```
[From]
????????????? ?.???? BTC
[To]
????????????? ?.???? BTC
????????????? ?.???? BTC
```

Sender
```
16aXLuZfnubmx
```

Of course, both proof needs to be zero-knowledge

Challenges

- **Proof to everyone:**
1) the transaction is valid; and
2) I submitted the required information to the regulator.

- **Efficiency of the proof**
- **Key management**

Project 4: Block chain protocols

- The fastest consensus protocol for permissioned block chain
  - >4500 TPS over 140 nodes @ < 40 sec. latency
  - >1300 TPS over 10K nodes @ < 40 sec. latency

- The first implementation of a DAG protocol (permission-less)
  - >1000 TPS over 10K nodes @ ~600 sec confirmation latency

Fintech: a big challenge for CS

- Instead of just application
- Calls for improvements to the CS core techniques

- AI and big data: unsupervised / semi supervised
- Distributed systems and consensus: how can we do it in the Internet
- Crypto techniques: can we have an efficient implementation?

- Wei Xu 徐葳 http://iiis.tsinghua.edu.cn/~weixu/