Uncertainty-Aware Event Analytics over Distributed Settings

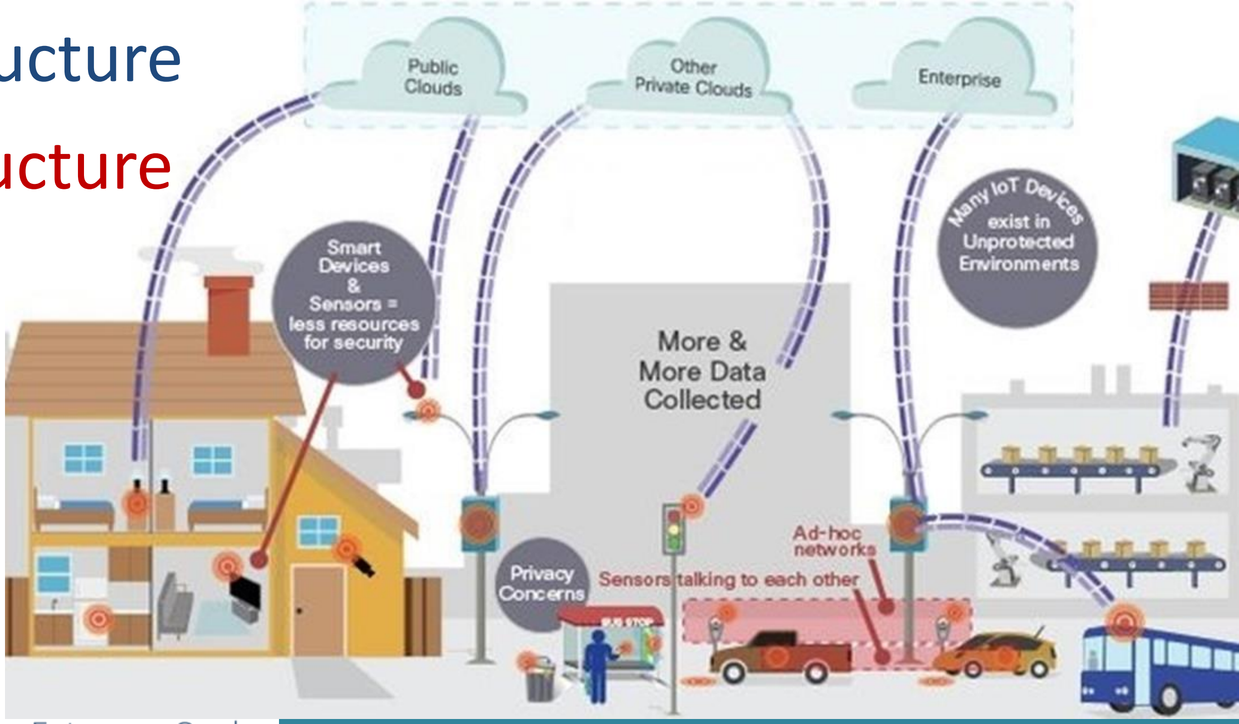
Nikos Giatrakos[§†], Alexander Artikis[*‡], Antonios Deligiannakis[§†], Minos Garofalakis[§†]

[§]Athena Research & Innovation Center, [*]University of Piraeus, [†]Technical University of Crete, [‡]NCSR Demokritos

INFORE
Interactive Extreme-Scale Analytics and Forecasting

(Geo)Distributed Architectures

- ~30 B connected devices by 2022 [Cisco VNI '18]
- Several data generation technologies
  - Smart Cities, Smart Grids, Smart Houses
  - Industry 4.0, Smart Factories
  - Telecom Infrastructure
  - Banking Infrastructure
  - Social Networks
  - Wearables
  - …



Big (Event) Data Challenges: 1-D B4 4-Vs

- Distribution: Massively distributed data streams →

  Need to reduce communication

- Volume

  NETWORK BOUND

- Velocity [e.g. Zleiter & Risch, PVLDB 2011, Karimov et al, ICDE 2018]
- Veracity (Uncertainty):
  - Imprecise Attribute Values, Uncertain Event Occurrence
  - Rules applied at a certain level of confidence
  - Event Forecasting, Approximation
- ~~Variety: various devices produce diverse data formats~~

E. Zeitler and T. Risch. Massive scale-out of expensive continuous queries. PVLDB, 4(11):1181–1188, 2011.
J. Karimov et al. Benchmarking Distributed Stream Data Processing Systems. ICDE, 1507-1518, 2018

Big (Event) Data Challenges: 1-D B4 4-Vs

- Distribution: Massively distributed data streams →
  Need to reduce communication
- Volume
- Velocity

NETWORK BOUND

[e.g. Zleiter & Risch, PVLDB 2011, Karimov et al, ICDE 2018]

- Veracity (Uncertainty):
  - Imprecise Attribute Values, Uncertain Event Occurrence
  - Rules applied at a certain level of confidence
  - Event Forecasting, Approximation

This Work: Handling Distribution + Uncertainty →

Boost manageable Volume and Velocity →

Extract Value (Event Analytics) out of Big Event Data

Our Contributions

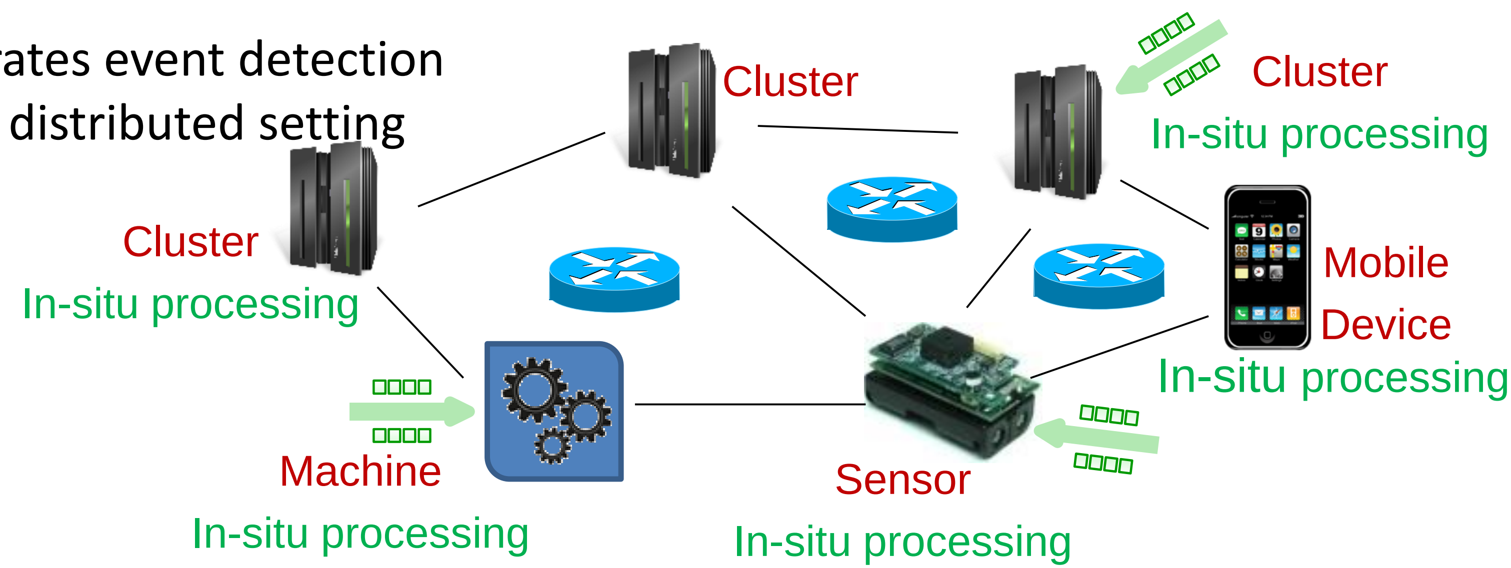
Generic Tools for Scalable Event-Analytics

- Tool 1: In-situ Processing
  - In-situ filter installation "safely" avoids communication
- Tool 2: Monitoring Protocol
  - Incorporates in-situ filters
  - Orchestrates event detection over the distributed setting

Integration in the FERARI Platform Prototype

I. Flouris et al. FERARI: A Prototype for Complex Event Processing over Streaming Multi-cloud Platforms. SIGMOD, 2093-2096, 2016.

I. Flouris et al. Complex event processing over streaming multi-cloud platforms: the FERARI approach. DEBS, 348-349, 2016



What Kind of Event-Analytics?

Event Data

- CEs: Complex Event Patterns
  - AGGRegation (Thresholded)
    - SUM, COUNT, AVG etc
    - lying above/below Threshold T
  - NON_AGGRegative Operator
    - AND: Logical Conjunction
    - OR: Logical Disjunction
    - SEQ: Time-ordered Conjunction
- (Un)Certainty/Confidence Threshold C
- SDEs: Simple Derived Events
  - Updates on $AGGR_j$

Target Queries/CE Detection

```
PATTERN NON_AGGR
(AGGR_1 > T_1,
 ...,
 AGGR_m > T_m) Q
[WHERE conditions]
[PARTITION BY key]
HAVING Q.Certainty> C
WITHIN window_const
```

Case Study: Mobile Fraud Detection

Q_1: FrequentToVoIPCalls

PATTERN(COUNT (CDR) > T) Q_1
WHERE CDR.prefix = VoIP
PARTITION BY CDR.callerID
HAVING Q_1.Certainty > C
WITHIN Y minutes

CDR = Call Detail Record

SDE Stream

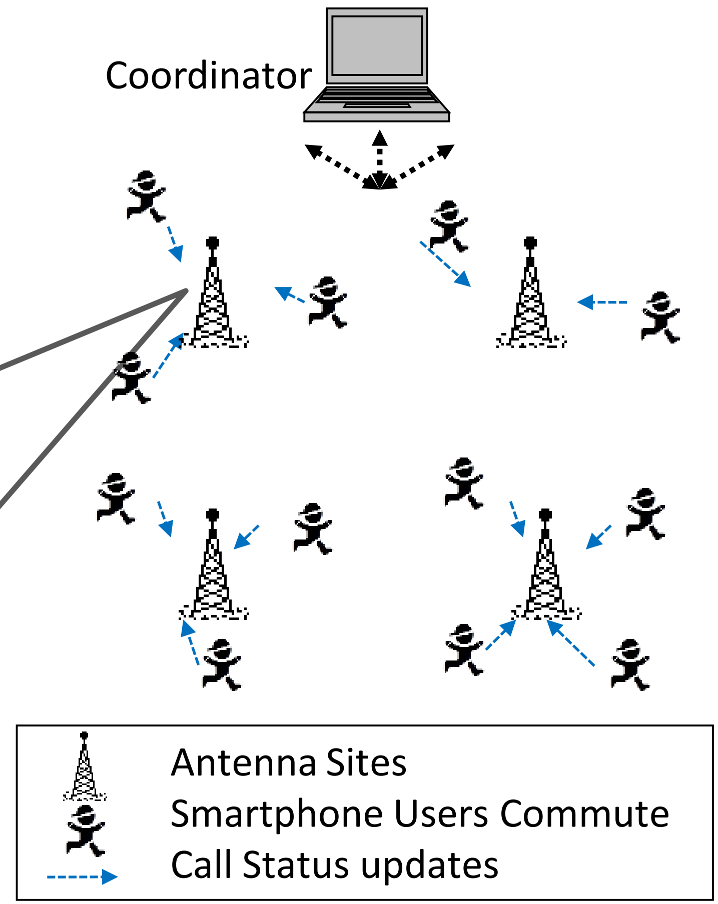
caller	callee	call start time	duration	p
62	23	11:10:23 05 - 10	22	0,41
38	45	11:10:24 05 - 10	21	0,43
34	22	11:10:23 05 - 10	13	0,41
83	19	11:10:25 05 - 10	6	0,42
10	22	11:10:24 05 – 10	6	0,4
34	41	11:10:24 05 - 10	9	0,41

Each VoIP call fraudulent with probability p

2-tiered Architecture

Coordinator – Query Source

N **sites - antennas**



Uncertainty-Aware In-situ Filters

Basic Concept: Suppress communication if no CEs can be produced

- Random Variable (R.V.) $X \equiv AGGR \in \{COUNT, SUM, \ldots\}$
- Global Filter @ Coordinator

$$1 - CDF[X, T] = P[X > T] \leq C$$

- In-situ Filters @ each site A_i (N antennas), R.V. $X_i \equiv AGGR_i$
  - If $X = \sum X_i \rightarrow$ $$CDF_i[X_i, T/N] \geq \sqrt[N]{1 - C}$$
  - If $X = \prod X_i \rightarrow$ $$CDF_i[X_i, \sqrt[N]{T}] \geq \sqrt[N]{1 - C}$$

Decomposable Probability Distributions

Distribution	PDF	Remarks	Decomposition Example	In-situ Filter for $1 - CDF(X,T) > C$
Normal	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$,	$\forall x \in \mathbb{R}$	$X_i \sim Normal(\mu_i, \sigma_i^2)$ $X = \sum_{i=1}^{N} X_i \sim Normal(\sum_{i=1}^{N} \mu_i, \sum_{i=1}^{N} \sigma_i^2)$	$\sqrt[N]{1-C} \le CDF(X_i, \frac{T}{N})$
Log-Normal	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$,	$\forall x > 0$ $\mu \in \mathbb{R}$(≠mean) $\sigma > 0$ (≠ st.dev.)	$X_i \sim LogNormal(\mu_i, \sigma_i^2)$ $X = \prod_{i=1}^{N} X_i \sim LogNormal(\sum_{i=1}^{N} \mu_i, \sum_{i=1}^{N} \sigma_i^2)$	$\sqrt[N]{1-C} \le CDF(X_i, \sqrt[N]{T})$
Chi-Square	$\frac{1}{2^{\nu/2}\Gamma(\frac{\nu}{2})}x^{\frac{\nu}{2}-1}e^{-\frac{x}{2}}$	$\forall x > 0$ $\nu \in \mathbb{N}^+$ degrees of freedom	$X_i \sim x^2(\nu_i)$ $X = \sum_{i=1}^{N} X_i \sim x^2(\sum_{i=1}^{N} \nu_i)$	$\sqrt[N]{1-C} \le CDF(X_i, \frac{T}{N})$
Cauchy	$\frac{1}{\pi s\left\lvert 1+(\frac{x-\nu}{s})^2\right\rvert}$	$\forall x \in \mathbb{R}$ $\nu \in \mathbb{R}$(location) $s > 0$ (scale)	$X_i \sim Cauchy(\nu_i, s_i)$ $X = \sum_{i=1}^{N} X_i \sim Cauchy(\sum_{i=1}^{N} \nu_i, \sum_{i=1}^{N} s_i)$	$\sqrt[N]{1-C} \le CDF(X_i, \frac{T}{N})$
Poisson	$\frac{\lambda^x e^{-\lambda}}{x!}$	$\forall x \in \mathbb{N}$ $\lambda > 0$	$X_i \sim Poisson(\lambda_i)$ $X = \sum_{i=1}^{N} X_i \sim Poisson(\sum_{i=1}^{N} \lambda_i)$	$\sqrt[N]{1-C} \le CDF(X_i, \frac{T}{N})$
Gamma	$\frac{1}{\Gamma(\alpha)\theta^{\alpha}}x^{\alpha-1}e^{-\frac{x}{\alpha}}$	$\forall x > 0$ $\alpha > 0$(shape) $\theta > 0$ (scale)	$X_i \sim Gamma(\alpha_i, \theta)$ $X = \sum_{i=1}^{N} X_i \sim Gamma(\sum_{i=1}^{N} \alpha_i, \theta)$	$\sqrt[N]{1-C} \le CDF(X_i, \frac{T}{N})$
Logistic	$\frac{e^{-\frac{x-\nu}{s}}}{s\left(1+e^{-\frac{x-\nu}{s}}\right)^2}$	$\forall x \in \mathbb{R}$ $\nu \in \mathbb{R}$(location) $s > 0$ (scale)	$X_i \sim Logistic(\nu_i, s_i)$ (approx.) $X = \sum_{i=1}^{N} X_i \sim Logistic(\sum_{i=1}^{N} \nu_i, \sqrt{\sum_{i=1}^{N} s_i^2})$	$\sqrt[N]{1-C} \le CDF(X_i, \frac{T}{N})$
Exponential	$\lambda e^{-\lambda x}$	$\forall x > 0$ $\lambda > 0$ (rate)	$X_i \sim Gamma(\frac{\alpha_i}{\sum_{i=1}^{N} \alpha_i}, \frac{1}{\lambda})$ $X = \sum_{i=1}^{N} X_i \sim Exp(\lambda)$	$\sqrt[N]{1-C} \le CDF(X_i, \frac{T}{N})$
Binomial	$\binom{n}{x}p^x(1-p)^{n-x}$	$x = 0, 1, \ldots, n$ $p \in [0, 1]$ $n \in \mathbb{N}$	$X_i \sim Binomial(n_i, p)$ $X = \sum_{i=1}^{N} X_i \sim Binomial(\sum_{i=1}^{N} n_i, p)$	$\sqrt[N]{1-C} \le CDF(X_i, \frac{T}{N})$

Case Study: Mobile Fraud Detection

Q_1: FrequentToVoIPCalls

```
PATTERN(COUNT (CDR) > T) Q1
WHERE CDR.prefix = VoIP
PARTITION BY CDR.callerID
HAVING Q1.Certainty > C
WITHIN Y minutes
```

CDR = Call Detail Record

- Each VoIP call fraudulent with probability $p \sim \mathrm{Bernoulli}[p]$
- n_i calls @ A_i, $n = \sum n_i$ total calls for a subscriber, $X = \sum X_i$
- $X_i \equiv COUNT_i \sim \mathrm{Binomial}[n_i, p] \rightarrow X \equiv COUNT \sim \mathrm{Binomial}[n, p]$
- Global Filter @ Coordinator

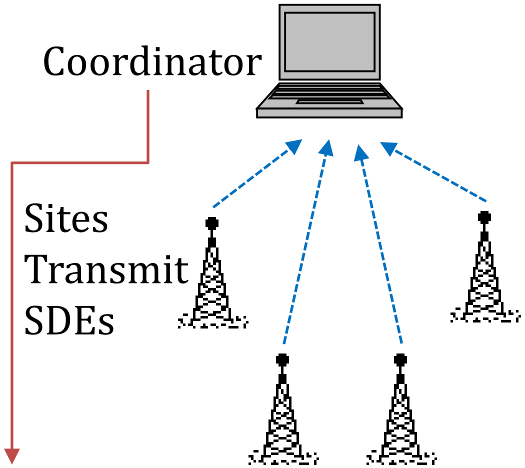
$$1 - CDF_{Binomial}[X, T] \leq C$$

- In-situ Filters @ each site A_i

$$CDF_{Binomial}[X_i, T/N] \geq \sqrt[N]{1 - C}$$

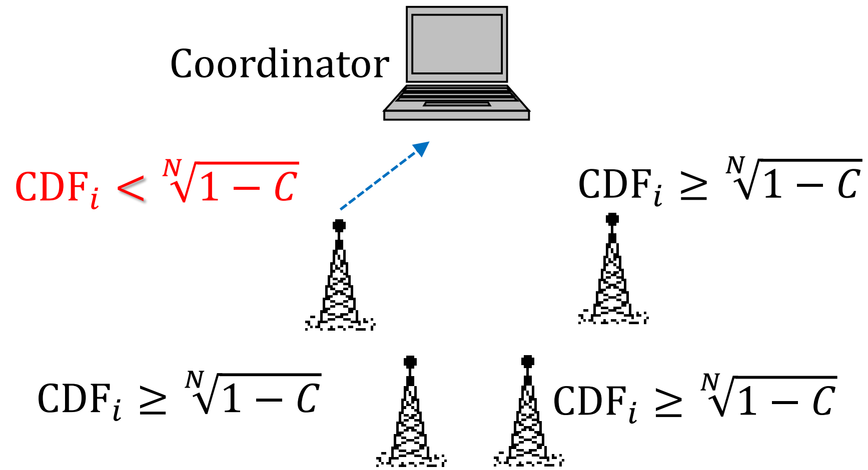
3-Phase Monitoring Protocol

Initialization Phase



1. Estimate PDF if not known in-hand
2. Set $X \sim PDF(.)$
3. Transmit $X_i \sim PDF(.)$ to each site A_i
4, Go to Monitoring Phase

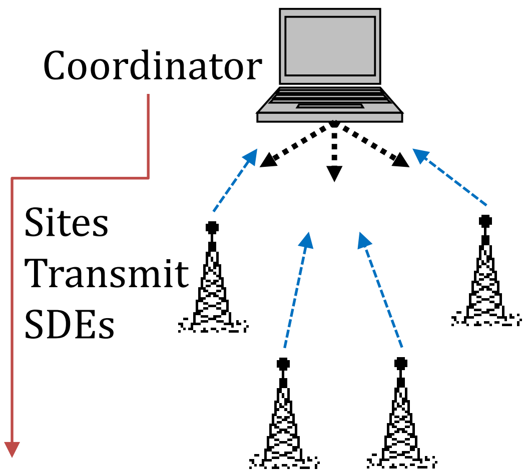
Monitoring Phase



$CDF_i \geq \sqrt[N]{1-C} \Rightarrow A_i$ caches relevant events
$CDF_i < \sqrt[N]{1-C} \Rightarrow A_i$ Synchronization Phase

3-Phase Monitoring Protocol

Synchronization Phase



Slack Allocation:

Adaptively increase or decrease the $\sqrt[N]{1-C}$ threshold for each site

1. Request cached events from sites $\{A_1, \cdots, A_N\}$

2.1 SyncCase A when $\Pr[X > T] > C$ [Global Filter violated] :

- 2.1.1 Produce CEs, receive new events
- 2.1.2 Go to 2.1
- 2.1.3 If $\Pr[X > T] \leq C$ [Global Filter holds]
  Go to Initialization phase

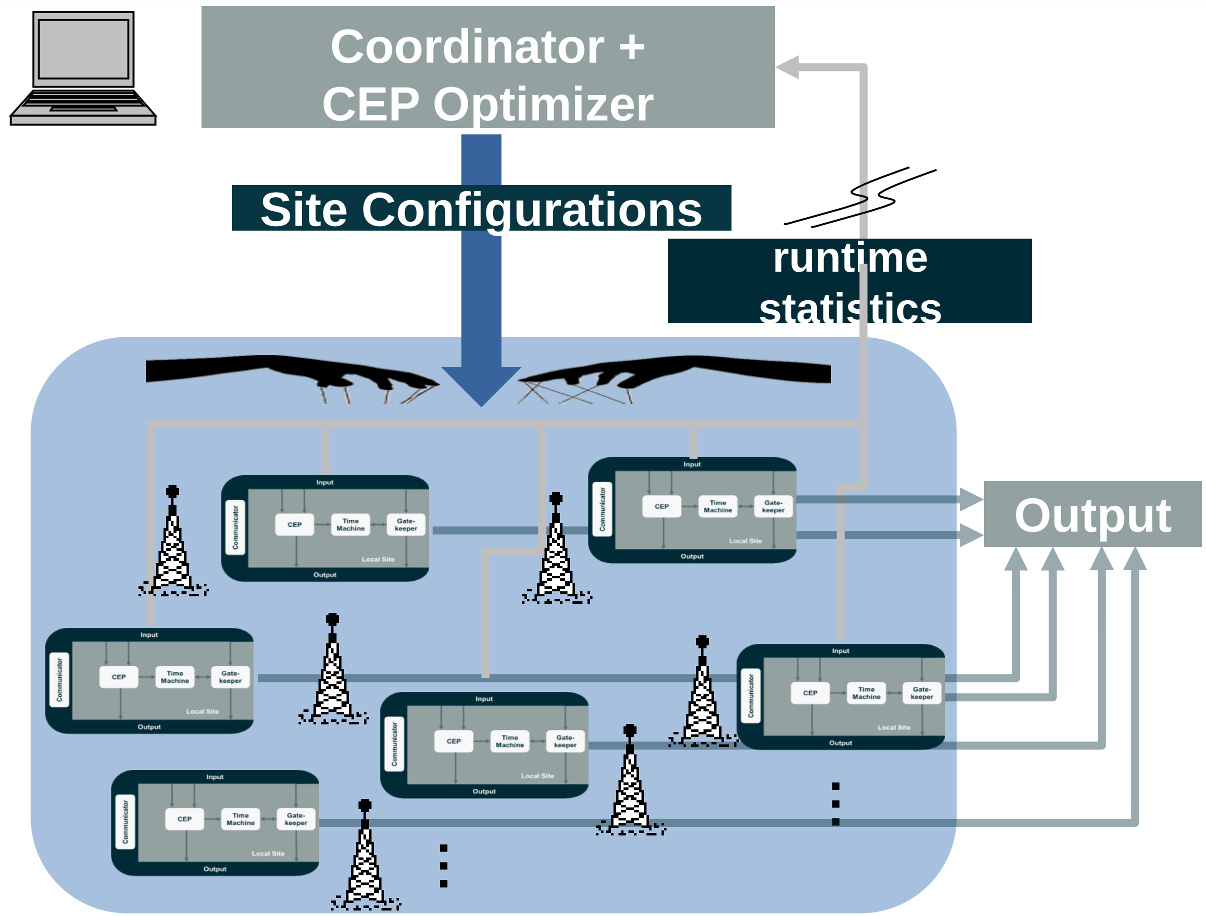
2.2 SyncCase B when $\Pr[X > T] \leq C$:

- 2.2.1 Slack Allocation
- 2.2.2 Go to Monitoring phase

Implementation in FERARI Platform

@ distributed architecture

FERARI Inter-site Orchestration



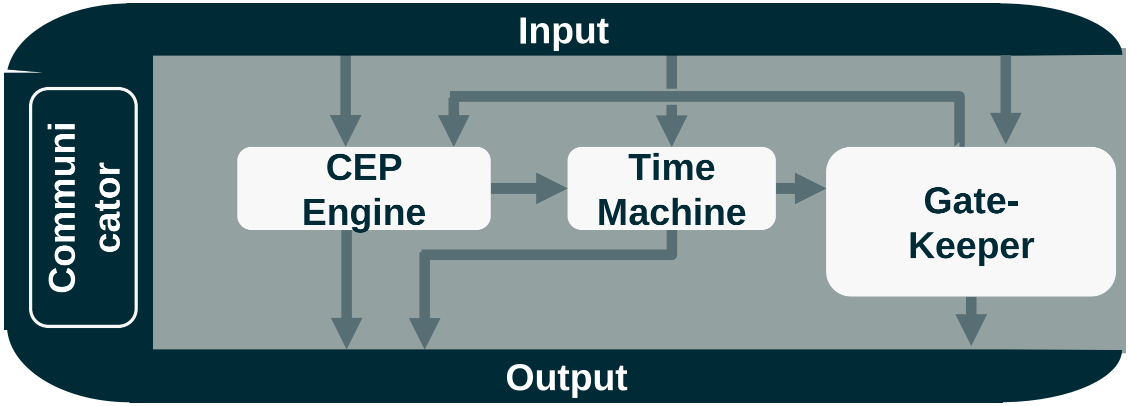
real-time input streams

I. Flouris et al. FERARI: A Prototype for Complex Event Processing over Streaming Multi-cloud Platforms. SIGMOD, 2093-2096, 2016.
I. Flouris et al. Complex event processing over streaming multi-cloud platforms: the FERARI approach. DEBS, 348-349, 2016

Implementation in FERARI Platform

@ each site

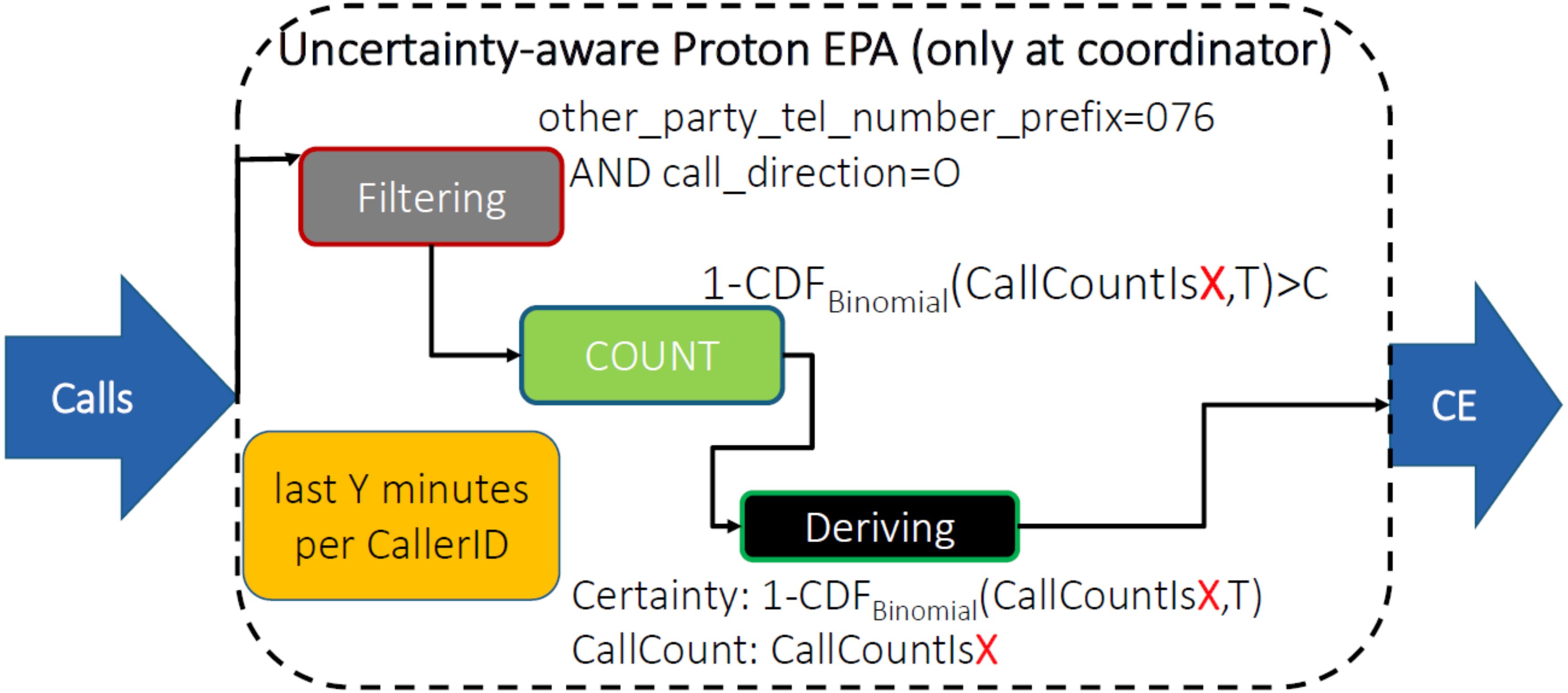
- Each site runs an Apache Storm topology
- Support any CEP Engine
- Current Implementations
  - ProtonOnStorm – IBM Haifa https://github.com/ishkin/Proton
  - Esper http://www.espertech.com/esper/
- Bridging the gap between two prototypes!



Traditional Implementation in Proton

- **Only @ coordinator** [Correia et al, DEBS 2015]
- No parallelism
- Naive central data collection at the coordinator

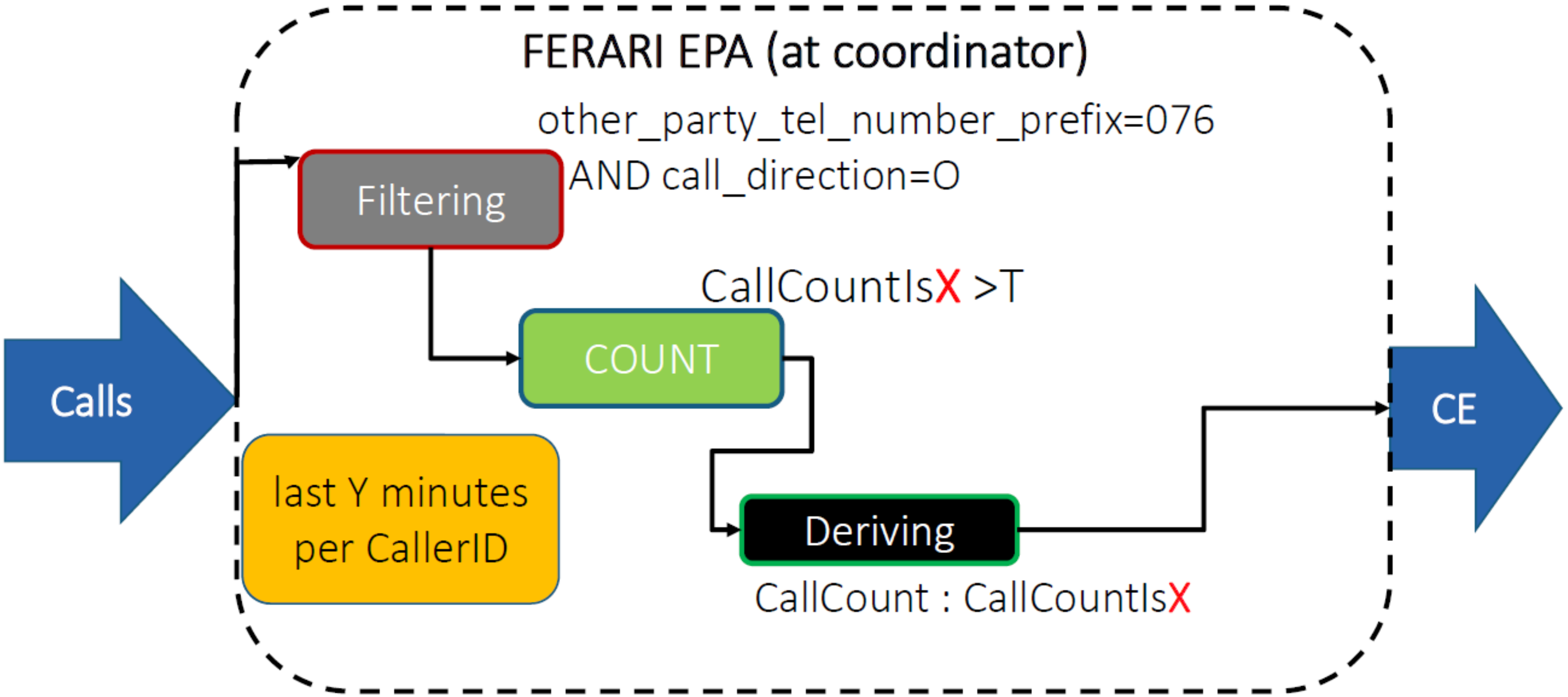
I. Correia et al. The uncertain case of credit card fraud detection. DEBS, 181-192, 2015



FERARI Implementation

@ coordinator

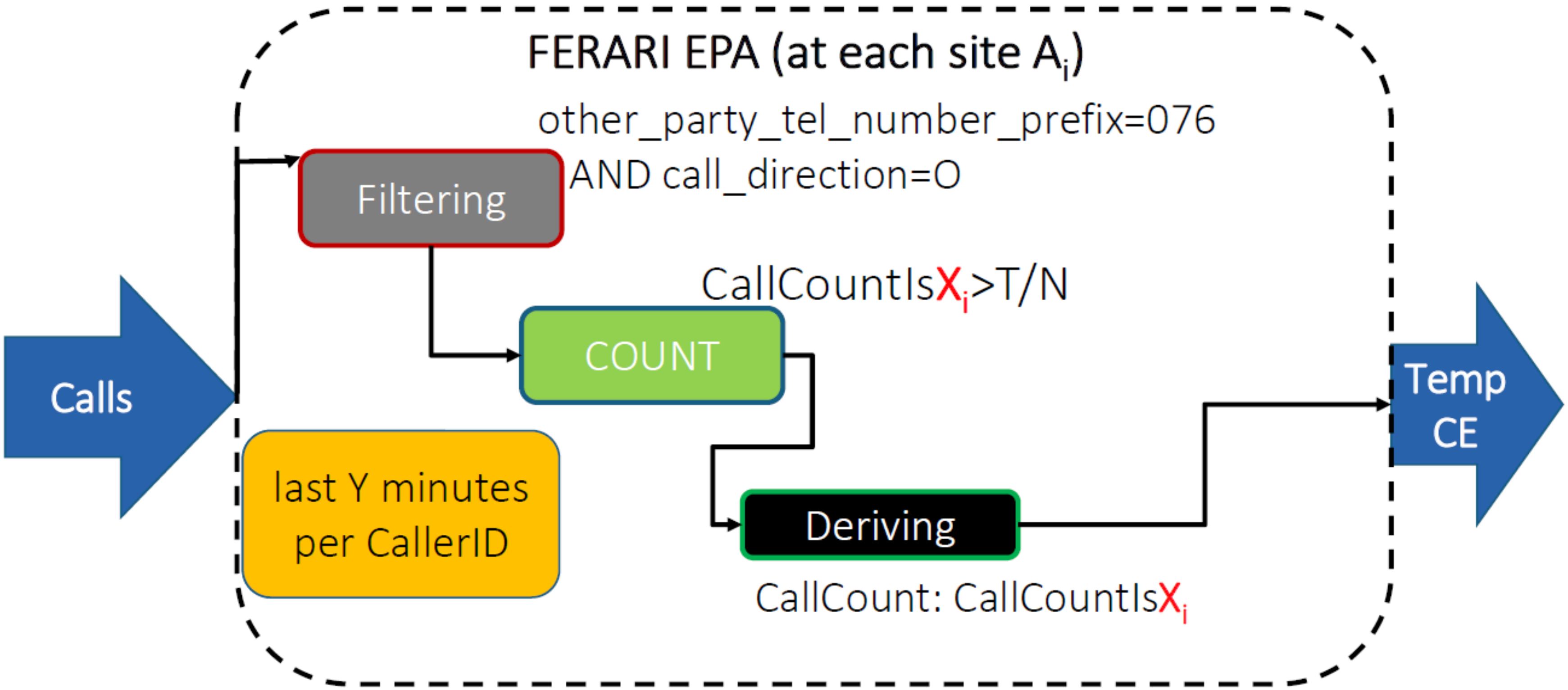
- Parallel processing in Apache Storm
- Monitoring protocol for network orchestration
- No support for uncertainty



FERARI Implementation

@ each site

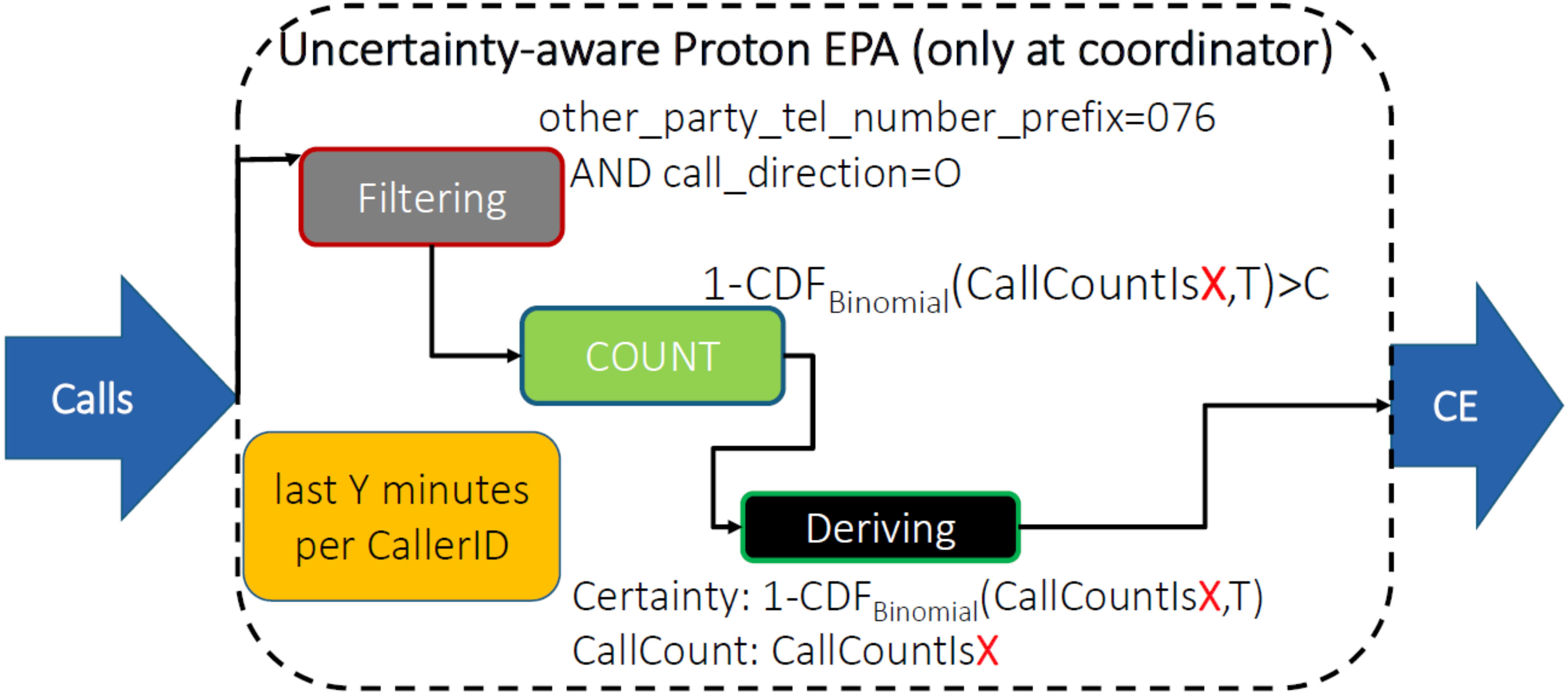
- Parallel processing in Apache Storm
- Monitoring protocol for network orchestration
- No support for uncertainty



This Work: Uncertainty-aware FERARI

@ coordinator

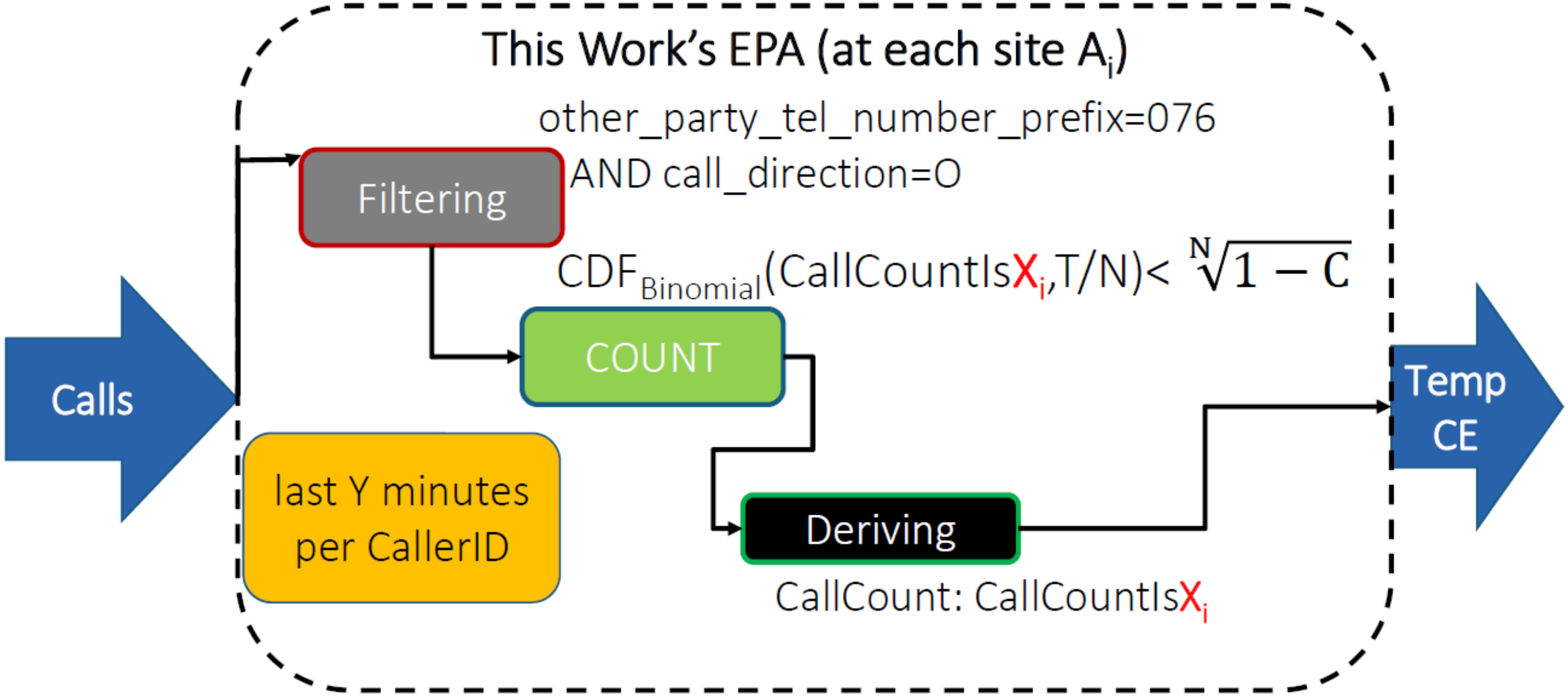
- Parallel processing in Apache Storm
- Monitoring protocol for network orchestration
- Support for uncertainty



This Work: Uncertainty-aware FERARI

@ each site

- Parallel processing in Apache Storm
- Monitoring protocol for network orchestration
- Support for uncertainty



Evaluation Results

Experimental Setup

- 160M calls from [Flouris et al, SIGMOD 2016]
- N=3 to N=10
- C=0.9 to 0.5
- Competitors
  - This Work
  - FERARI + Uncertainty-Aware Coordinator
  - Naïve central data collection (omitted)

Highlights

- N=3, C=0,9 → An order of magnitude less transmitted messages
- On average 4 times less transmitted messages across various N and C
- N → 10 or C → 0.5 no earnings
  - Recall: $CDF_i[X_i, T/N] \geq \sqrt[N]{1-C}$
  - As N increases
    - $\sqrt[N]{1-C} \rightarrow 1$
    - $T/N \rightarrow 0$

Summary & Future Work

- Optimized distributed execution of uncertainty-aware event queries
- Communication Reduction
  - Construction and installation of In-situ Filters at sites
- Network Orchestration
  - Introduction of monitoring protocol
- Proof-of-Concept
  - Extending FERARI streaming multi-cloud platform
- Real case study from the telecom domain
- Future work:
  - Sampling among sites to increase performance
  - Loosen the uncertainty independence assumption

http://infore-project.eu/

Uncertainty-Aware Event Analytics over Distributed Settings

Thank you!
Questions?

Nikos Giatrakos, Alexander Artikis,
Antonios Deligiannakis, Minos Garofalakis